Thierry Hatt

Contribution à

Hatt-Diener, Marie-Noèl, « Strasbourg à la croisée des chemins, étude des mobilités urbaines , 1810-1840 », 2 tomes, Université de Tours, 2000, 502 et 528 p.

INFORMATIQUE ET RECHERCHE HISTORIQUE

I. Matériels, logiciels, fichiers

Ce travail était impossible sans recourir à l'informatique. Une base de données de 500 000 champs, 14 000 enregistrements, 6 400 personnes et 3 900 maisons ne peut-être traitée à la main. Le recours à l'informatique dans toutes ses composantes a donc été d'emblée un choix stratégique essentiel de la recherche[1]. Le travail a été réalisé sur une longue période de temps si l'on considère qu'il a commencé avec la base de données de l'enquête de 1789 qui a été réalisée de 1988 à 1989 – soit une durée de plus dix ans -.

Pour l'évolution du logiciel cela représente une durée immense et des modifications importantes ; nous avons, par exemple, dû nous adapter à trois changements majeurs de systèmes d'exploitation disponibles sur les machines utilisées, des « compatibles IBM PC » (passage de DOS aux premières versions de Windows puis à Windows 95). Cela a aussi re^présenté le passage de l'interface ardue du Dos à la convivialité graphique de Windows.

Les progrès informatiques ont été considérables pendant cette période. L'un des progrès décisifs, par exemple, a été la facilité accrue des tris croisés[2] dont nous avons fait abondamment usage. Dans la phase la plus ancienne du travail, si le chercheur ne pouvait avoir recours aux logiciels poids lourds des centres de calcul (SAS, SPCSS, …) il devait lui-même écrire les logiciels de tri croisé[3], ce n'est plus le cas aujourd'hui.

---

[1] On trouvera une bonne introduction dans J. L. Pinol, A. Zysberg, « Métier d'historien avec l'ordinateur », Nathan Paris, 239 p. 1995. Cet ouvrage s'arrête en deçà de l'analyse des données.

[2] Rappelons à propos du tri croisé qu'il est très peu pratiqué avant l'ère informatique étant donnée sa grande lourdeur de mise en œuvre.

[3] On trouvait ce type de programmes élémentaires de dépouillement d'enquêtes dans F. Pottier, « Initiation à l'informatique dans les sciences de l'homme », Classiques Hachette, Paris, 153 p, 1973 ; ou bien dans L. Lebart, J. P. Fénelon, « Statistique et informatiques appliquées », Dunod Paris, 457 p. 1973. C'était souvent en Fortran, il fallait donc convertir dans le langage disponible, L. S. E. puis Pascal dans notre cas pour la version Turbo Sesam. On en trouvait encore édités en 1985 sous forme de programmes Basic : M. Roux, « Algorithmes de classification », Masson, Paris, 152 p. 1985. L'apport des Anglo Saxons a été

Un autre progrès majeur est l'amélioration de la qualité de publication, en particulier en ce qui concerne les sorties graphiques grâce aux imprimantes laser et jet d'encre couleur bon marché. Ces progrès n'ont pas résolu tous les problèmes, loin de là, on verra quelques-unes des difficultés posées par l'informatique individuelle dans une telle recherche.

Un tel travail nécessite de recourir à une panoplie étendue de logiciels dont la maîtrise est loin d'être aisée et prend du temps. La chaîne de production complète qui va des données à la publication est composée de logiciels autonomes, les échanges entre eux sont parfois malaisés. Il paraît difficile que l'historien domine à la fois sa discipline et les outils techniques nécessaires. Des soutiens techniques sont très utiles.

Nous présenterons d'abord l'environnement matériel et logiciel dans lequel nous avons travaillé. Sans chercher à dissimuler les problème rencontrés, nous donnerons ensuite des exemples complets de démarches techniques : une procédure de requête de base de données pour obtenir une reconstitution de famille ; quelques indications précises à propos de la méthode de calcul des écarts à l'indépendance de Michel Volle dont nous avons assuré la programmation, les étapes de construction d'une carte géographique, combinaison de l'usage de la base de données, du tableur et de cartographeur, les étapes complètes de réalisation d'un analyse des données.

A. LES MATERIELS : AISANCE DES STOCKAGES VITESSE DES TRAITEMENTS

Les ordinateurs utilisés à la fin de la période de travail (comme les logiciels ils ont changé plusieurs fois), sont trois Pentium II ou III en réseau local, une machine serveur sous NT, les autres postes sous Windows 95 et Windows 98 ; 50 Go sur neuf disques en ligne sur le réseau, accessibles pour tous les utilisateurs, un graveur de CD-ROM. On est bien loin des disquettes 5 pouces 1/4 de 360 Ko de capacité des années 1989.

B. LES LOGICIELS : BELLE DISPONIBILITE, RICHESSE DES FONCTIONS

Une belle gamme de logiciels puissants sont maintenant disponibles. Reste au chercheur à les prendre en main.

---

aussi essentiel on citera par exemple : M. R. Anderberg, « Cluster analysis for applications », Academic Press, New York, 359 p. 1973 ou bien R. S. Baxter « Computer and statistical techniques for planners », Methuen, London, 336 p. 1976.

1. Bases de données

Notre choix initial s'est porté sur Foxbase – clone de Dbase III- sous DOS. Ce logiciel a été conservé pour effectuer la saisie, sous Windows, son fonctionnement a été très correct. Les traitements proprement dits, listes, tris simples, tris croisés, jointures … on été réalisés avec deux versions successives d'Access pour Windows, la dernière étant Access 2000.

2. Langages de programmation

Ils étaient obligatoires en 1988-89 pour effectuer des tris croisés, indisponibles en standard sous Dbase ou Foxbase. Mais Access réalise les tris croisés avec une simplicité merveilleuse et cette étape de programmation est donc dépassée. Il a quand même fallu programmer des « filtres » dans la phase plus récente pour les conversions de format de fichiers et pour mettre en œuvre la méthode des écarts à l'indépendance de Michel Volle. Nous avons d'autre part utilisé le langage « quatrième génération » de Dbase ainsi que Turbo-Pascal dans ses versions 4 à 7 et Maple[4] V. 4 pour publier la méthode Volle.

3. Logiciels d'« analyse des données »[5]

Dans la phase la plus ancienne de la recherche, il existait très peu de logiciel de ce type ou des produits très lourds disponibles uniquement sur gros systèmes, SAS, SPCS …,, nous avions donc été amenés à rédiger TURBO-SESAM[6] sous Pascal. Ce logiciel a encore servi pour la phase la plus récente de l'étude du corpus puis nous l'avons progressivement remplacé par des produits sous Windows : Stat-Lab de Michel Jambu[7] (versions 2 et 2.1a) – ce produit semble ne plus exister - et

---

[4] http://www.maplesoft.com

[5] Il s'agit des méthodes d'analyse multidimensionnelles, connues sous le nom d'analyse factorielle et d'analyse hiérarchique. On trouvera des éléments essentiels sur ce type de démarche dans l'ouvrage d'un maître français J. P. Benzécri, « L'analyse des données », t I, « la taxinomie », t II, « l'analyse des correspondances », Dunod, Paris, 615 et 619 p. 1978.

[6] Thierry Hatt, : « SESAM : système d'étude spatiales et d'analyses multidimensionnelles, manuel d'utilisation », INRDP, SIE, 69 p, 1977.

« Sesam : manuel d'utilisation II, les programmes d'analyse factorielle des correspondances et de cartographie automatique », INRP, 1979.

« Sesam V. II », logiciel LSE (disquettes et mode d'emploi de 50 p), CNDP, 5/1983.

« Sesam V. II : Système d'études spatiales et d'analyse multidimensionnelle, manuel », IREM, ULP Strasbourg, 158 p, 6/1983.

La version en Turbo-Pascal de SESAM date de 1986-1987

[7] On se reportera aux travaux de M. Jambu, « Classification automatique pour l'analyse des données », Dunod, Paris tome I et II, 310 et 399 p. 1978. L. Lebart, A. Morineau, J. P. Fénelon, « Traitement des données statistiques, méthodes et programmes », Dunod, Paris, 512 p, 1979. Eux aussi proposaient de nombreux programmes en Fortran.

Spad de Alain Morineau versions 3 et 4[8] de la CISIA. Ce dernier produit est disponible aussi bien sur PC que sur Apple, il est de grande qualité.

4. Logiciels de cartographie

Le travail portant sur 1789 avait été entièrement cartographié sous TURBO-CARTO[9], produit que nous avions rédigé spécialement à une époque où ce type de logiciel était soit financièrement inabordable soit même inexistant. Nous avons remplacé ce produit par des logiciels sous Windows : Didger de Golden Software pour la numérisation des fonds de carte ; pour le système d'information géographique multi-couches d'abord MapInfo version DOS puis Mapviewer version 3 de Golden Software[10].

5. Logiciels grapheurs

Il nous semble nécessaire de recourir à des produits spécialisés. Les graphiques disponibles sous Excel 4.0 ne répondent pas suffisamment aux critères de la sémiologie graphique. Nous avons utilisé Graphwin et Grapher (V. 2) version Dos et Windows de Golden Software.

6. Logiciel d'imagerie

Nous avons eu recours à ces logiciels pour certains fonds de carte complexes, pour la numérisation des registres et la présentations des cartes anciennes de Strasbourg ainsi que pour les captures d'écran de Spad, l'exportation de fichiers images ne fonctionnant pas.correctement : Paint Shop Pro versions 4 et 5 sous Windows, Foto Look pour la numérisations des images et un numériseur de table Agfa.

7. Logiciel tableur

Le tableur s'est révélé fondamental en particulier comme interface d'import-export de données entre les différents logiciels[11]. Nous avons utilisé Excel 3 puis 4 enfin 2000. Un grave inconvénient de ce produit est son incapacité à reconnaître les dates antérieures à 1900 ce qui nous a obligé à faire des acrobaties – en particulier des conversions fréquentes de type caractère / date pour passer d'Access à Excel.

---

[8] http://www.cisia.com/

[9] Thierry Hatt, modules cartographiques de Turbo-Sesam.

[10] http://www.goldensoftware.com/

[11] Voir le schéma de circulation de l'information : Le cheminement de l'information dans le traitement de la base de données

8. Logiciels de publication et de traitement de texte

La saisie des textes a été faite avec deux versions de Word sous Windows, la dernière étant Word 2000 sans problème particulier de compatibilité. Il n'est pas certain que ce soit le meilleur choix pour un travail de plus neuf-cent pages abondamment fourni en annexes graphiques. Le montage final s'est en particulier révélé laborieusement lent et parfois pénible. Il a fallu par exemple réduire fortement la résolution des images pour cause de débordement mémoire (la machine support du traitement de texte Word disposait pourtant de 128 Mo de RAM …). Le traitement des fichiers ASCII avec Ultra EDIT 32 6.20b a été nécessaire en particulier pour préparer les fichiers issus de l'analyse des données.

9. Logiciels d'indexation documentaire de gestion des fichiers et de gravure de CD-ROM

Nous avons eu recours à ce produit gratuit disponible sur l'Internet pour indexer l'ensemble des documents par le contenu : Altavista Discovery V1.1.
Logiciel FileSync 2.14 a été utilisé pour les sauvegardes incrémentales des fichiers d'un ordinateur du réseau à l'autre en tenant compte des dates de dernière mise à jour. Nous avons utilisé CRWin 3.6A de Golden Hawk et CD Creator 3.5 d'Adaptec pour graver les CD-ROM, les déchets ont été rares, liés plutôt à Windows (profondeur des arborescences trop grande).

On peut remarquer le poids de l'investissement pour un chercheur solitaire qui ne serait pas soutenu par une institution. L'ensemble de ces logiciels représente plusieurs milliers de francs. Même si la dépense est ici étalée sur plusieurs années, cela reste fort important, d'autant plus que les « mises à jour », obligées, sont très souvent facturées au prix fort.

C. UNE GRANDE VARIETE NECESSAIRE : LES FORMATS DE FICHIERS

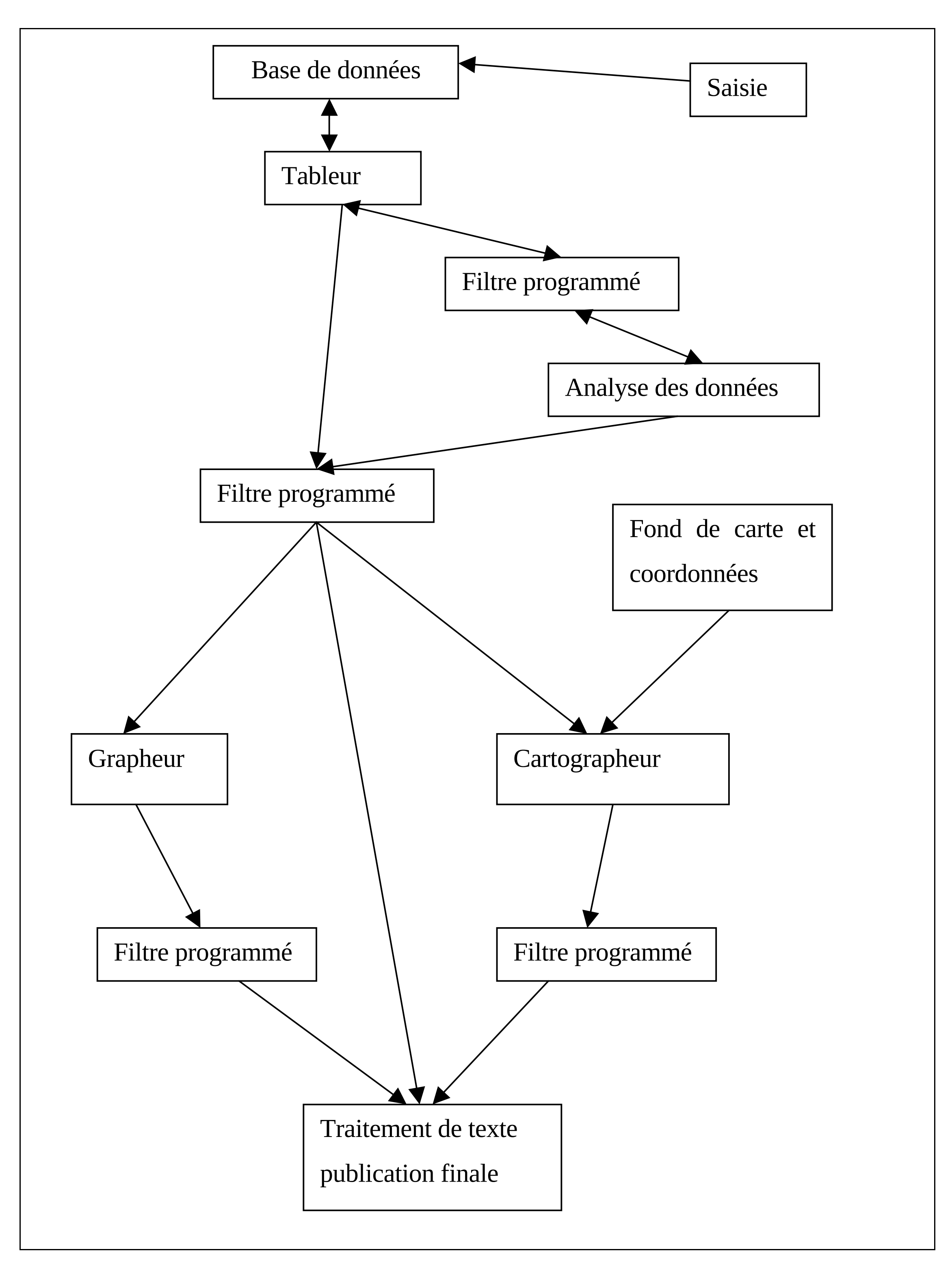
L'un des principaux problèmes est la nécessité de faire passer les données et les résultats de chaque logiciel vers les autres. La situation n'est pas aussi facile que la norme Windows généralisée pourrait le laisser penser. Dans de nombreux cas le seul format possible d'échange est le format ASCII associé à un traitement. Ce traitement peut-être simple : nettoyage avec un traitement de texte de caractères empêchant la relecture mais il peut aussi être complexe et nécessiter une programmation spécifique. Ces programmes ont été écrits en Dbase ou Turbo-Pascal. Ces programmes spécifiques appelés « filtres programmés » sur le graphique du cheminement de

l'information[12] ont été en particulier nécessaires pour la transmission des fonds de carte vectoriels d'un logiciel à l'autre.

1. Près d'une vingtaine de formats de fichiers différents ont été utilisés :

- Bases de données :Formats DBF pour Foxbase et MDB pour Access, format Ascii texte ou xls pour les échanges de données entre les logiciels de bases de données et d'analyse de données, entre Access et Excel.
- Logiciels d'analyse des données : TURBO-SESAM (format Ascii), Spad, Statlab import de fichiers textes ou xls.
- Logiciels grapheurs : Graphwin; Grapher (formats Ascii ou Excel et Wmf)
- Logiciels de cartographie : TURBO-CARTO, MapInfo, SIG Mapviewer, format DXF ou Wmf, Ascii ou formats propriétaires (GSB, GSM et BNA pour les outils Mapviewer), export en WMF vectoriel pour l'intégration au traitement de texte Word 2000.
- Logiciels d'imagerie : Paint Shop Pro (format jpg pour l'Internet, gif pour le traitement de texte, tif pour la numérisation).
- Logiciels tableurs format ASCII et XLS.
- Logiciel de développement de site Internet : formats html, ASCII, .doc ou .rtf
- Logiciel de traitement de texte : format .doc de Word ; les échanges des logiciels de traitement vers le traitement de texte ont été faits en mode jpg, gif pour les images pixels et wmf pour les images vectorielles.

[12] Voir en page 7



2. Le cheminement de l'information dans le traitement de la base de données jusqu'à la publication finale

D. LES FICHIERS DE L'ETUDE : UN GRAND NOMBRE NECESSAIRE

1. *Variété et volume des données*

Dans une situation typique (en fin de travail) le volume de fichiers disponibles en ligne a atteint plus de 600 Mo, plus de 1000 fichiers répartis de la façon suivante sans compter les fichiers de la base de données de recherche contextuelle (0.9 Go) :

LES DONNEES		
TYPE D'APPLICATION	**NOMBRE DE FICHIERS**	**VOLUME DES DONNEES**
Analyse des données	300	90 Mo
Bases de données	40	130 Mo
Cartographie	540	260 Mo
Tableur	300	90 Mo
Grapheur	130	10 Mo
Traitement de texte	70	40 Mo
Ensemble des fichiers de données du travail	Soit 1000~fichiers	~600 Mo
Altavista Discovery	300	872 Mo
En tout	Soit ~1300 fichiers	Soit ~1,4 Go de données
LES LOGICIELS D'APPLICATION		
Ensemble des logiciels listés (base de données, tableau, grapheur etc…)	Plusieurs milliers de fichiers	1,2 Go

Le volume total de fichiers manipulés pendant la durée complète de l'étude a été en réalité plus important, certaines données plus rarement utilisées (par exemple les images des fonds de carte pour l'Europe ou bien les numérisations des registres) n'étant pas gardées en ligne mais sur CD-ROM.

2. *Une politique de sauvegarde rigoureuse facilitée par les supports disponibles*

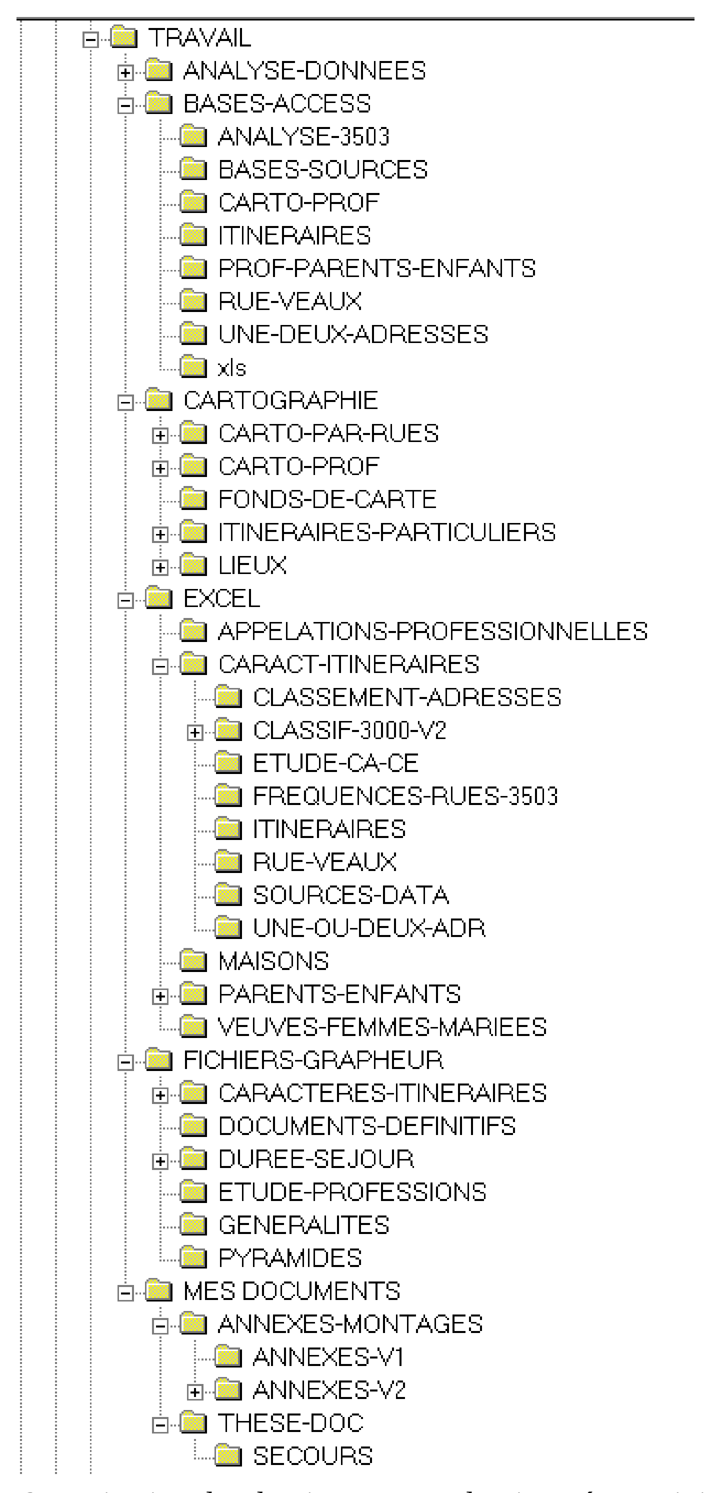
La sauvegarde a été assurée sur CD-ROM, leur prix en août 2000 a baissé jusqu'à 6 francs ce qui a permis de multiplier les sécurités. Une politique très rigoureuse de sauvegarde a été mise en place.

Les données ont été systématiquement sauvegardées sur trois ordinateurs différents en réseau, en mode incrémental. Les versions successives de travaux aux étapes t, t+1, t+2 étant disponibles à tout moment. Dès que les travaux le nécessitaient un CD-ROM était gravé, stocké de temps en temps dans un lieu physiquement différent de la salle informatique pour éviter une perte ou une destruction simultanée des originaux et des copies. Une trentaine de Cd-ROM ont été gravés soit 13 Go de données, données dans lesquelles il fallait à l'occasion chercher des états anciens mais nécessaires.

3. Quelques méthodes de gestion des données

Ce nombre et ce volume de fichiers ne peuvent plus être gérés « à la main ». L'une des principales difficultés étant de retrouver les informations et les « bons » fichiers. Nous avons donc utilisé Altavista Discovery pour indexer cette masse d'information par le contenu et retrouver les fichiers dans les cas les plus difficiles. Très efficace, cette indexation ne prenait en compte que les données en ligne, pas les données sur CD-ROM. Il fallait donc retrouver ces derniers « à la main ».

Le principe de classement retenu sur le disque dur est le suivant : les données sont stockées sur une unité physique différente du système d'exploitation, sur cette unité, chaque logiciel spécialisé, base de données, traitement de texte, grapheur …, possède son propre répertoire, à l'intérieur de ce répertoire les sous dossiers sont répartis par thème de la recherche historique : les migrants, l'analyse de l'échantillon des 3 502 personnes, l'étude de la rue des Veaux, les mobilités professionnelles …. Une centaine de répertoires différents ont été créés au fur et à mesure des besoins.



Organisation des dossiers et sous dossiers, état en juillet 2000

E. QUELQUES DIFFICULTES A SURMONTER

1. La faible qualité des publications informatiques sur le marché rend parfois difficile certains apprentissages

A notre grande surprise, nous étant portés acquéreurs d'ouvrages permettant d'aller « plus loin » avec Access ou Word, nous avons eu le désagrément de constater que, très souvent, malgré leur prix élevé, ces livres ne sont que des répliques de l'aide en ligne ou des collections de recettes très difficilement transposables ; ils ne permettent donc pas de compléter l'aide du logiciel très souvent insuffisante. Nous n'avons, par exemple, pas trouvé de descriptif complet à jour des fonctions disponibles dans Access ce qui a rendu l'usage du logiciel pénible ,pour le calcul des durées, les conversions de date en chaîne, l'extraction de sous-chaîne dans les champs de la base... Ces difficultés sont aggravées par une très grande difficulté à trouver de l'aide du côté des utilisateurs expérimentés.

2. Difficulté pour trouver de l'aide auprès d'autres utilisateurs

La plupart des logiciels utilisés dans ce travail sont complexes et paramétrables à l'infini (Spad 4.0, par exemple, nécessite pour une analyse factorielle en composantes principales suivie d'une classification hiérarchique la fixation de 60 paramètres, certains sont définis par défaut, mais tous doivent être passés en revue et vérifiés). La plupart des gens n'utilisent qu'une toute petite partie des possibilités de ce type de logiciel. Il est alors très rare de trouver une personne qui, ayant rencontré le même problème que vous est à même de vous aider.
Soit le problème n'est pas traité de la même façon, soit la version du logiciel n'est pas la même, soit le problème n'a pas été rencontré sous cette forme. Autrement dit, le chercheur qui ne bénéficie pas d'un laboratoire bien équipé en personnel technique, et disposant des mêmes logiciels, se retrouve seul devant son informatique. Cette difficulté est aggravée par les lacunes et défauts des produits.

3. Lacunes et défauts des logiciels

Nous ne donnerons que quelques exemples.

L'incapacité d'Excel de reconnaître les dates antérieures à 1900

Cette lacune du logiciel nous a obligé à des manipulations compliquées pour le passage d'Access à Excel et pour le calcul des dates. Elles ont été converties en caractères, non sans peine : la fonction de conversion n'est renseignée correctement nulle part ! Il a donc fallu tâtonner. Les calculs sur les

dates devaient tous être faits sous Access pour être ensuite exportés vers Excel pour la réalisation d'histogramme ou d'autres traitements. Ces constants changements de type des variables de l'étude ont été une source constante de problèmes d'autant plus que les messages d'erreur associés sont souvent peu explicites. Nous avons néanmoins préféré cette solution au décalage de deux siècles. Plus compliquée, certes elle permet d'assurer des calculs plus précis.

La complexité de mise en œuvre de certaines opérations

Quelques exemples : le calcul d'histogramme avec Excel ; sous Word, la gestion du « document maître », le référencement des notes de bas de page ou le positionnement correct du numéro de page ; sous Access, l'ajout de lignes à une base de données, la création d'une nouvelle table à partir de données éparses sont complexes et difficiles à reproduire sans un grand entraînement.

Les plantages

Ils peuvent occasionner des pertes de données, même si une sauvegarde périodique est prévue. A cet égard Word s'est révélé particulièrement fragile et a parsemé le disque de fichiers temporaires non effacés par suite de sorties prématurées et catastrophiques du logiciel. Aucun logiciel n'est épargné et Spad, par exemple, a posé de gros problèmes à l'importation des données texte et ODBC.

Le manque de clarté ou l'absence de messages d'erreur est un défaut de la plupart des logiciels.

C'est un facteur de ralentissement du travail. Il est très difficile de corriger une erreur dont la localisation n'est pas du tout précisée alors que la requête fautive compte plusieurs dizaines d'instructions ! La liste des erreurs graves est longue : les numéros de page qui n'apparaissent plus, les cartes en couleur qui sont imprimées en une seule couleur, le plan qui n'apparaît plus après un changement apparemment minime, nous aurions pu composer un bêtisier complet, sauf quand l'erreur entraîne un plantage et une perte du système.

Les erreurs de relecture des fichiers d'un logiciel à l'autre

Versions de format non reconnus, images importées non orthonormées ou totalement dégradées par rapport à l'original, trop grandes ou trop petites, fichiers vectoriels importés à des échelles erronées ou impossibles à modifier, fichiers Excel bien relus mais impossible à visualiser à l'écran à cause d'un format de cellule mal relu, impossibilité du copier-coller alors que les deux logiciels sont tous deux « pour Windows », mystérieuses erreurs de relecture des données Excel par Access, erreurs

soigneusement collectées dans une base Access mais pas plus compréhensibles pour autant, l'incapacité d'un certain logiciel de relire les lettres« oe » liées à la mode Windows et empêchant à cause de ce simple caractère la lecture complète d'un tableau de données….Des formats d'image vectoriels bien exportés par un logiciel et mal importé par un autre, là encore la liste est longue.

La variété des problèmes liés aux changements de format est infinie et pourtant, c'est le seul moyen d'échanger des données d'un logiciel à l'autre. Toutes ces difficultés obligent à des manœuvres complexes de contournement qui font perdre énormément de temps.

Problèmes liés au vaste volume des données et à la variété des questions à traiter

Une nécessaire gestion rigoureuse des fichiers

Ce travail a duré plusieurs années. Les outils, les données, les démarches, les méthodes d'approche ont donc évolué avec le temps. Il fallait périodiquement revenir sur des travaux faits plusieurs mois auparavant sur des données incomplètes ou bien comprises autrement. Ces retours en arrière doivent être géré avec soin.

Plusieurs méthodes ont été choisies : la tenue d'un journal de bord des fichiers a vite trouvé ses limites devant l'explosion du nombre, une autre méthode a consisté à dater systématiquement les requêtes dans le fichier lui-même, une autre enfin à stocker les données de manière structurée, une autre à faire figurer sur les imprimés systématiquement le nom du fichier d'origine, une dernière à utiliser Altavista Discovery.

La datation est le moyen de mémoriser les tâches effectuées qui a le mieux réussi, il était assez aisé de se rappeler quels traitements avaient été faits à telle ou telle époque. Cette datation doit évidemment être attachée au fichier, la date de dernière utilisation du système d'exploitation ne peut suffire puisque le simple fait d'ouvrir ce fichier modifie la date. Le deuxième moyen, déjà décrit, a consisté à sauvegarder les données par thème. Cette méthode a eu ses limites pour plusieurs raisons : la nécessité des échanges d'un logiciel à l'autre a obligé à stocker certaines données Excel dans la zone « base de données » ou dans la zone « grapheur » au lieu de l'espace « tableur » et a donc obligé à rechercher en plusieurs endroits au lieu d'un seul ; d'autre part certaines requêtes pouvaient avoir été traitées sous Excel ou bien sous Access encore fallait il s'en souvenir ! La dernière méthode, d'indexation par le contenu sous Altavista Discovery s'est révélé très efficace, elle a permis de retrouver très rapidement un fichier par son contenu, encore fallait-il se rappeler des mots clefs présents dans le dit fichier.

Toutes ces méthodes de mémorisation ne nous ont pas empêché parfois d'être obligés de reconstruire complètement une procédure faute de retrouver les données source associées.

Variété des demandes exploratoires à satisfaire

La recherche sur un sujet qui n'a pas été exploré exige la construction « cousue-main » de procédures et de démarches propres qui peuvent difficilement être récupérées ailleurs ou être « industrialisées ».C'est une difficulté mais aussi un des intérêts d'un tel travail. Certaines approches techniques sont bien connues ; tri croisé, jointure, élimination des doublons, cartographie en cercles proportionnels, requêtes logiques mais leur adaptation aux données est nécessaire, une bonne connaissance du logiciel doit s'accompagner d'une très bonne connaissance des données. Dans le cas contraire, selon la formule des Anglo-Saxons : « garbage in, garbage out », si vous rentrez des données fantaisistes dans le logiciel il vous renverra des résultats juste dignes d'être mis à la « poubelles ». Cette collaboration étroite entre le spécialiste du corpus et l'informaticien est indispensable. Elle est difficile car l'historien sait bien ce qu'il veut mais ne sait pas comment l'obtenir, l'informaticien sait comment procéder mais ne comprenant pas forcément toutes les subtilités du corpus il peut se lancer dans des traitements dont les résultats sont erronés.

Une fois que la procédure informatique de traitement a été mise au point, ce qui peut-être long, l'exécution est faite en quelques secondes mais le résultat n'est pas garanti car l'erreur de paramétrage est la règle, pas l'exception ! Il est donc absolument impératif que le chercheur spécialiste de son dossier passe les résultats au crible pour les rejeter éventuellement et revoir les procédures.

L'« industrialisation » de la « production »

Chaque question posée aux données par le chercheur étant spécifique il est difficile d'industrialiser la production. C'est évidemment ennuyeux car l'informatique donne sa pleine puissance dans ce cas. L'intérêt de l'informatique est à chercher ailleurs, dans la souplesse infinie des traitements possibles, dans le volume énorme des données manipulables. Il est pourtant possible de temps en temps d'optimiser la démarche de production lorsqu'une collection de documents de même type est demandée – tableaux, cartes, histogrammes –.

F. NECESSITE IMPERIEUSE D’APPARTENANCE A UNE EQUIPE OU A UN RESEAU

Comme on peut le voir dans ce panorama rapide les problèmes à résoudre sont nombreux, parfois très techniques. Ces questions ne peuvent que difficilement être résolues par une personne seule. L’appartenance à un laboratoire, à une équipe de recherche tenant des séminaires réguliers, à une association nationale ou internationale, le soutien de techniciens voire d’ingénieurs informatiques sont nécessaires. Il est indispensable qu’il y ait des lieux où les questions techniques puissent être posées et ne pas faire croire que le chercheur est à même de résoudre toutes ces questions. Malheureusement l’évolution récente vers la micro-informatique a laissé dans de nombreux cas le chercheur seul face à son ordinateur individuel alors même que, progrès matériels et logiciels aidant, les exigences de « production » sont de plus en plus élevées. Il faut être conscient que cette situation est difficile à assumer.

II. Procédures statistiques et cartographiques

A. EXPLOITATION DE LA BASE DE DONNEES DES DECLARATIONS EXEMPLE DE PROCEDURE COMPLEXE, ETUDE DES RELATIONS PARENTS ENFANTS

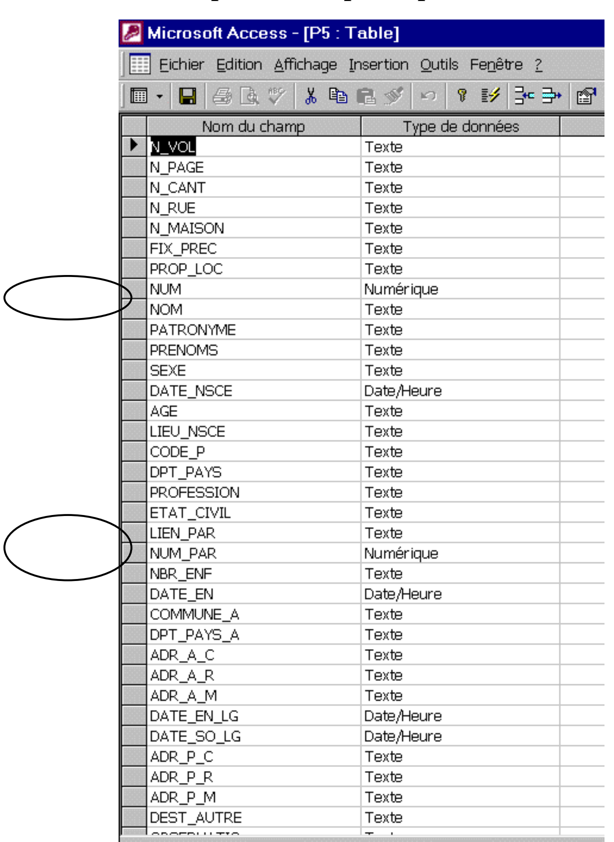
1. La source des données

Deux bases de données sont utilisées dans cet exemple de traitement : la base source des 13844 déclarations et une base annexe : la base des filiations, liste des associations parents enfants.

P5 : base de données source des 13844 déclarations

Liste parents enfants : sous base des filiations

P5 : Structure de la base de données source : Le champ NUM désigne l'identificateur unique de l'individu étudié. Le lien de parenté de ce dernier LIEN_PAR le caractérise et NUM_PAR pointe vers le parent identifié. Tous les champs ne sont pas représentés ici.

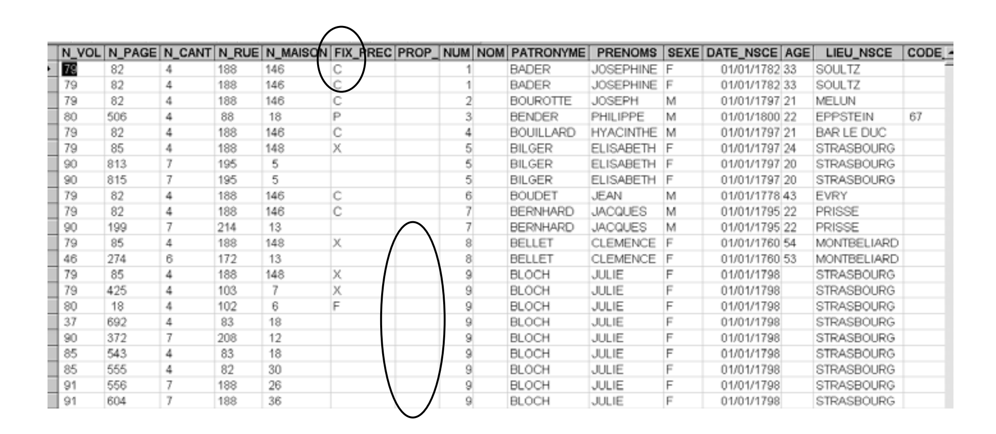


Microsoft Access - [P5 : Table]

Fichier Edition Affichage Insertion Outils Fenêtre ?

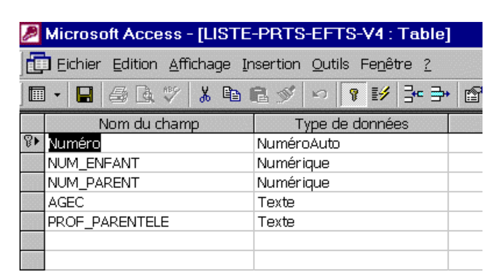
Nom du champ	Type de données
N_VOL	Texte
N_PAGE	Texte
N_CANT	Texte
N_RUE	Texte
N_MAISON	Texte
FIX_PREC	Texte
PROP_LOC	Texte
NUM	Numérique
NOM	Texte
PATRONYME	Texte
PRENOMS	Texte
SEXE	Texte
DATE_NSCE	Date/Heure
AGE	Texte
LIEU_NSCE	Texte
CODE_P	Texte
DPT_PAYS	Texte
PROFESSION	Texte
ETAT_CIVIL	Texte
LIEN_PAR	Texte
NUM_PAR	Numérique
NBR_ENF	Texte
DATE_EN	Date/Heure
COMMUNE_A	Texte
DPT_PAYS_A	Texte
ADR_A_C	Texte
ADR_A_R	Texte
ADR_A_M	Texte
DATE_EN_LG	Date/Heure
DATE_SO_LG	Date/Heure
ADR_P_C	Texte
ADR_P_R	Texte
ADR_P_M	Texte
DEST_AUTRE	Texte

P5 : début de la liste des données ; chaque personne est enregistrée autant de fois qu'elle a fait une déclaration de déménagement avec toutes les informations disponibles au moment de cet acte. Ici les neuf déclarations de la personne portant (c'est un hasard) le NUM 9.



N_VOL	N_PAGE	N_CANT	N_RUE	N_MAISON	FIX_PREC	PROP_	NUM	NOM	PATRONYME	PRENOMS	SEXE	DATE_NSCE	AGE	LIEU_NSCE	CODE_
79	82	4	188	146	C		1		BADER	JOSEPHINE	F	01/01/1782	33	SOULTZ	
79	82	4	188	146	C		1		BADER	JOSEPHINE	F	01/01/1782	33	SOULTZ	
79	82	4	188	146	C		2		BOUROTTE	JOSEPH	M	01/01/1797	21	MELUN	
80	506	4	88	18	P		3		BENDER	PHILIPPE	M	01/01/1800	22	EPPSTEIN	67
79	82	4	188	146	C		4		BOUILLARD	HYACINTHE	M	01/01/1797	21	BAR LE DUC	
79	85	4	188	148	X		5		BILGER	ELISABETH	F	01/01/1797	24	STRASBOURG	
90	813	7	195	5			5		BILGER	ELISABETH	F	01/01/1797	20	STRASBOURG	
90	815	7	195	5			5		BILGER	ELISABETH	F	01/01/1797	20	STRASBOURG	
79	82	4	188	146	C		6		BOUDET	JEAN	M	01/01/1778	43	EVRY	
79	82	4	188	146	C		7		BERNHARD	JACQUES	M	01/01/1795	22	PRISSE	
90	199	7	214	13			7		BERNHARD	JACQUES	M	01/01/1795	22	PRISSE	
79	85	4	188	148	X		8		BELLET	CLEMENCE	F	01/01/1760	54	MONTBELIARD	
46	274	6	172	13			8		BELLET	CLEMENCE	F	01/01/1760	53	MONTBELIARD	
79	85	4	188	148	X		9		BLOCH	JULIE	F	01/01/1798		STRASBOURG	
79	425	4	103	7	X		9		BLOCH	JULIE	F	01/01/1798		STRASBOURG	
80	18	4	102	6	F		9		BLOCH	JULIE	F	01/01/1798		STRASBOURG	
37	692	4	83	18			9		BLOCH	JULIE	F	01/01/1798		STRASBOURG	
90	372	7	208	12			9		BLOCH	JULIE	F	01/01/1798		STRASBOURG	
85	543	4	83	18			9		BLOCH	JULIE	F	01/01/1798		STRASBOURG	
85	555	4	82	30			9		BLOCH	JULIE	F	01/01/1798		STRASBOURG	
91	556	7	188	26			9		BLOCH	JULIE	F	01/01/1798		STRASBOURG	
91	604	7	188	36			9		BLOCH	JULIE	F	01/01/1798		STRASBOURG	

Deuxième base de données : structure de la sous base de filiation. Ces informations ont été saisies séparément. A chaque NUM d'enfant est associé un NUM de parent, un âge calculé et la profession de la parentèle.

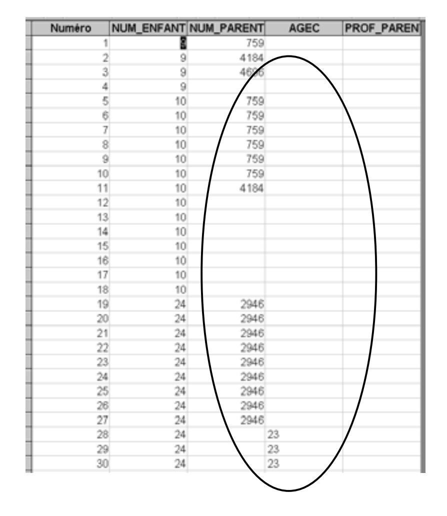


Microsoft Access - [LISTE-PRTS-EFTS-V4 : Table]

Fichier Edition Affichage Insertion Outils Fenêtre ?

Nom du champ	Type de données
Numéro	NuméroAuto
NUM_ENFANT	Numérique
NUM_PARENT	Numérique
AGEC	Texte
PROF_PARENTELE	Texte

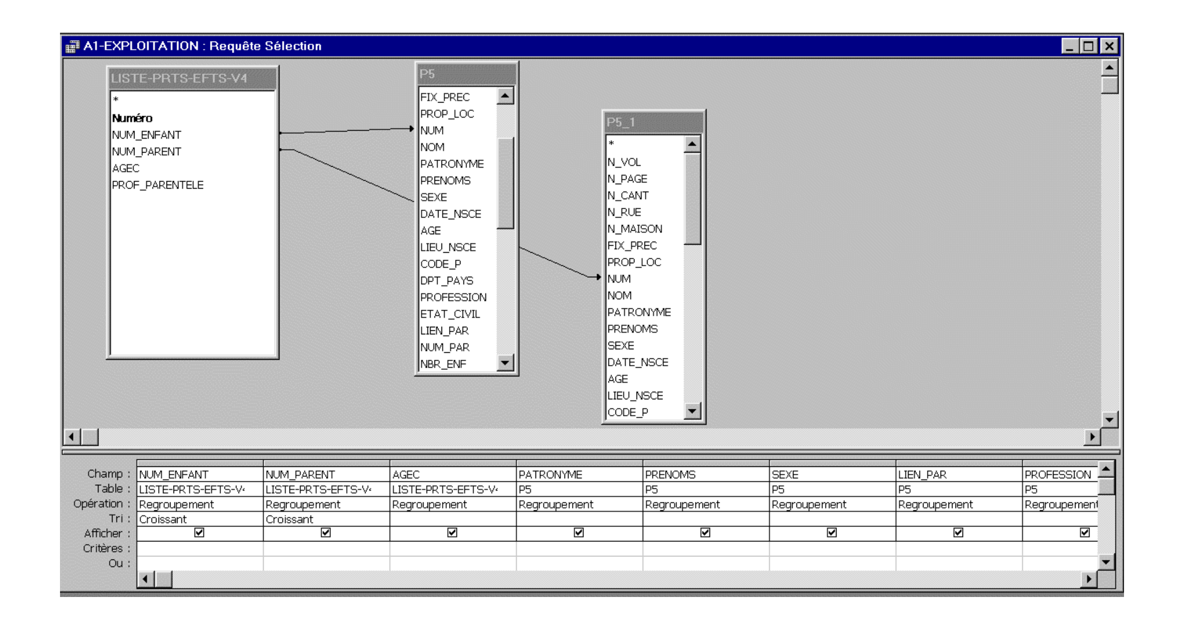
Liste de filiation, les données : certains champs peuvent être vides



Numéro	NUM_ENFANT	NUM_PARENT	AGEC	PROF_PAREN
1	9	759		
2	9	4184		
3	9	4696		
4	9			
5	10	759		
6	10	759		
7	10	759		
8	10	759		
9	10	759		
10	10	759		
11	10	4184		
12	10			
13	10			
14	10			
15	10			
16	10			
17	10			
18	10			
19	24	2946		
20	24	2946		
21	24	2946		
22	24	2946		
23	24	2946		
24	24	2946		
25	24	2946		
26	24	2946		
27	24	2946		
28	24		23	
29	24		23	
30	24		23	

2. *La reconstitution des familles*

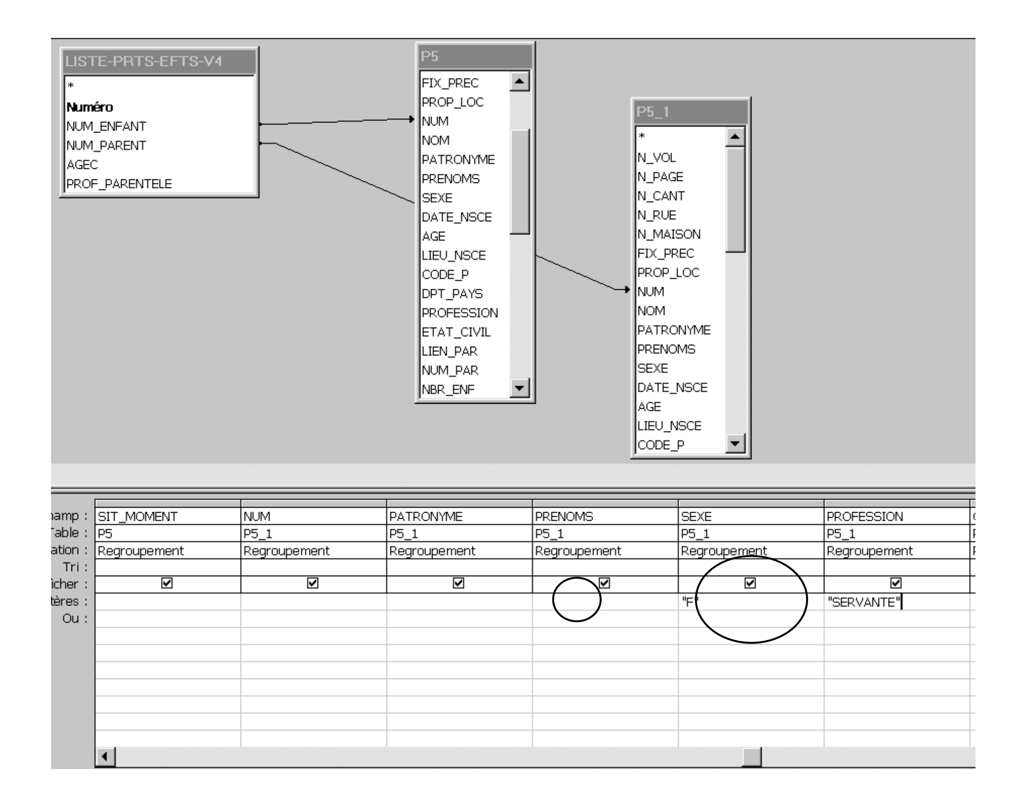
Création à partir de la liste de filiation et par « jointure [13]» avec la base complète d'une sous base des enfants et de leurs parents, le champ NUM enfant ne peut pas être clef primaire sinon on n'aurait pas tous les parents. L'opération utilisée ici est une jointure « externe gauche », suivie d'un « regroupement » pour avoir toutes les situations différentes et seulement celles ci. Pour améliorer la présentation des données tri sur le NUM enfant.



---

[13] La jointure est l'opération qui consiste à lier plusieurs bases en utilisant une « clef » commune. La jointure est dite « externe gauche » =i. e. tous les enregistrements du côté gauche sont ajoutés aux résultats de la requête, même si le champ joint de la table située à droite ne contient pas de valeurs correspondantes ; par contre les enregistrements de la table de droite ne sont combinés à ceux de la table de gauche que si les champs joints comportent des valeurs correspondantes. Ceci permet d'obtenir les parents et leurs enfants.

A partir de ce résultats création de la sous base des « mères dont la profession est servante » : le sexe du parent est « féminin » et la profession « servante ».



Exemple de résultat : professions des enfants de mères qui ont déclaré être servante au moins une fois :

NUM_ENFANT	P5_1.SEXE	ETAT_CIVIL	P5_1.PROFESSION	NUM_PARENT	P5.SEXE	P5.PROFESSION
101	F		SERVANTE	5899	F	SERVANTE
101	F	CE	SERVANTE	5899	F	SERVANTE
146	F		COUTURIERE	2852	F	
146	F		COUTURIERE	2852	F	FILLE PUBLIQUE
146	F		COUTURIERE	2852	F	OUVRIERE
146	F		COUTURIERE	2852	F	SERVANTE
146	F	FM		2852	F	
146	F	FM		2852	F	FILLE PUBLIQUE
146	F	FM		2852	F	OUVRIERE
146	F	FM		2852	F	SERVANTE
146	F	FM	COUTURIERE	2852	F	
146	F	FM	COUTURIERE	2852	F	FILLE PUBLIQUE
146	F	FM	COUTURIERE	2852	F	OUVRIERE
146	F	FM	COUTURIERE	2852	F	SERVANTE
146	F	FM	JOURNALIERE	2852	F	
146	F	FM	JOURNALIERE	2852	F	FILLE PUBLIQUE
146	F	FM	JOURNALIERE	2852	F	OUVRIERE
146	F	FM	JOURNALIERE	2852	F	SERVANTE
146	F	FM	OUVRIERE AUX TABACS	2852	F	
146	F	FM	OUVRIERE AUX TABACS	2852	F	FILLE PUBLIQUE
146	F	FM	OUVRIERE AUX TABACS	2852	F	OUVRIERE
146	F	FM	OUVRIERE AUX TABACS	2852	F	SERVANTE
596	F	CE	SERVANTE	561	F	SERVANTE
650	F			632	F	
650	F			632	F	COUTURIERE

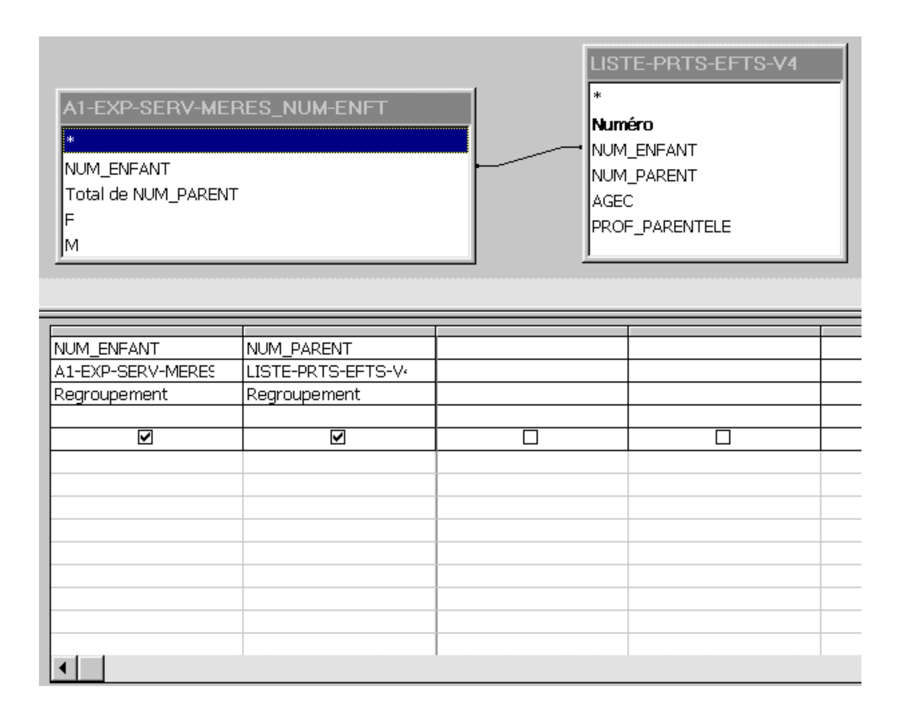
Exploitation de cette sous base : élimination des doublons des enfants par tri croisé des champs NUM enfant x SEXE

NUM_ENFANT	Total de NUM_	F	M
4625	4		4
2387	1		1
3030	5		5
4302	4		4
4570	2		2
6135	2		2
5538	1		1
6057	1		1
6104	2		2
596	1	1	
1089	1	1	
1117	1	1	
1778	1	1	
5899	1	1	
5564	1	1	
5566	1	1	
4787	1	1	
5565	1	1	
101	2	2	
4738	2	2	
5413	2	2	
650	2	2	
5721	3	3	
5226	3	3	
3546	3	3	
146	8	8	

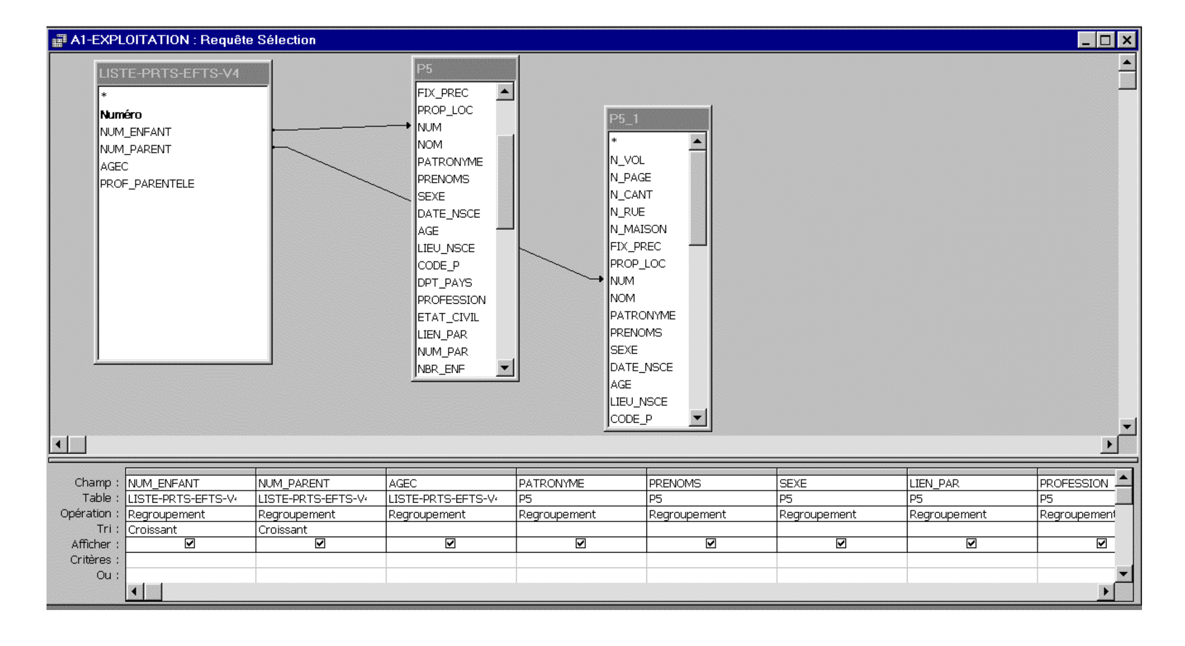
Même opération pour avoir la liste des NUM des parents ; croisement du NUM et du code professionnel

NUM_PARENT	Total de NUM_	SE
101	3	3
167	3	3
308	1	1
554	2	2
561	1	1
632	2	2
1089	1	1
1117	1	1
1176	1	1
1452	5	5
2141	2	2
2388	1	1
2852	10	10
4119	3	3
4773	1	1
4976	10	10
5325	4	4
5504	2	2
5721	1	1
5899	2	2

A partir du champ NUM unique de l'enfant (dont la mère est servante) on pointe sur la liste de filiation pour obtenir tous les NUM de parents différents :



Recherche des variantes professionnelles des mères servantes (au moins une fois) ; jointure entre la liste de filiation et deux copies de la base principale :



3. *Liste des professions des enfants et des parents*

NUM_ENFANT	NUM_PARENT	P5.PATRONYME	P5.PROFESSION	NUM	P5_1.PATRONYME	SEXE	P5_1.PROFESSION
101	5899	BEYERLE	SERVANTE	5899	BEYERLE	F	SERVANTE
101	5899	BEYERLE	SERVANTE	5899	BEYERLE	F	SERVANTE
146	2852	BARTHELEMY		2852	BARTHELEMY	F	SERVANTE
146	2852	BARTHELEMY	COUTURIERE	2852	BARTHELEMY	F	SERVANTE
146	2852	BARTHELEMY	COUTURIERE	2852	BARTHELEMY	F	SERVANTE
146	2852	BARTHELEMY	COUTURIERE	2852	BARTHELEMY	F	SERVANTE
146	2852	BARTHELEMY	JOURNALIERE	2852	BARTHELEMY	F	SERVANTE
146	2852	BARTHELEMY	OUVRIERE AUX TABACS	2852	BARTHELEMY	F	SERVANTE
146	2852	BARTHELEMY	COUTURIERE	2852	BARTHELEMY	F	SERVANTE
146	2852	BARTHELEMY	COUTURIERE	2852	BARTHELEMY	F	SERVANTE
596	561	BLESSIG	SERVANTE	561	BLESSIG	F	SERVANTE
650	632	BERLICHY		632	BERLICHY	F	SERVANTE
650	632	BERLICHY		632	BERLICHY	F	SERVANTE
1089	1117	BALLADE	SERVANTE	1117	BALLADE	F	SERVANTE
1117	1089	BALLADE	SERVANTE	1089	BALLADE	F	SERVANTE
1778	1176	BADER		1176	BASTARD	F	SERVANTE

B. « UNE METHODE POUR LIRE ET COMMENTER AUTOMATIQUEMENT DES GRANDS TABLEAUX STATISTIQUES »

Michel Volle, a donné cette méthode d'analyse dans Economie et Statistique, Insee, n° 52, p. 46-50, 1974. Nous avons adapté et programmé en L.S.E[14]., en Pascal puis en Maple la procédure proposée par Michel Volle en 1974 et republiée en 1997 dans « Analyse des données » chez Economica. Nous décrivons ici la méthode qui a été utilisée pour traiter les tableaux croisés de la thèse.

1. *Méthode et calculs*

On considère un tableau t d'effectifs, donc de valeurs positives, à nli lignes et ncol colonnes. On commence par calculer les marges en ligne fl (somme de tous les éléments en colonne) les marges en colonne fc (somme de tous les éléments en ligne) respectivement (après initialisation à zéro des tableaux fl et fc) :

$$fc_j = \left(\sum_{i=1}^{nli} t_{i,j}, j = 1 .. ncol \right) \qquad fl_i = \left(\sum_{j=1}^{ncol} t_{i,j}, i = 1 .. nli \right)$$

[14] Thierry Hatt, « Sesam : Système d'Etudes Spatiales et d'Analyse Multidimensionnelle » CNDP, 1982.

On calcule le total général du tableau :

$$tg = \sum_{j=1}^{ncol} \left(\sum_{i=1}^{nli} t_{i,j} \right)$$

On calcule ensuite les fréquences marginales, et les fréquences du tableau en divisant tous les éléments par le total général :

$$fc_j = \left(\frac{fc_j}{tg}, j = 1 .. ncol \right) \qquad fl_i = \left(\frac{fl_i}{tg}, i = 1 .. nli \right)$$

$$f_{i,j} = \left(\frac{t_{i,j}}{tg}, i = 1 .. nli, j = 1 .. ncol \right)$$

« Les marges du tableau fc et fl apportent, à elles seules, une certaine information sur la population étudiée. » Il est évident que le tableau complet en apporte davantage.

Connaissant les marges du tableau on peut construire un tableau fictif « produit des marges », prod où lignes et colonnes sont proportionnelles entre elles à caractères indépendants.

$$prod_{i,j} = fl_i \, fc_j, \; i = 1 .. nli, j = 1 .. ncol$$

La méthode propose de comparer le tableau réel des fréquences avec le tableau fictif « produit des marges ». Le gain d'information que l'on fait en passant du tableau fictif au tableau réel est calculé ainsi :

$$Info = \sum_{j=1}^{ncol} \left(\sum_{i=1}^{nli} \frac{(f_{i,j} - fl_i \, fc_j)^2}{fl_i \, fc_j} \right)$$

En connaissant uniquement les marges on possède une certaine information sur le tableau, en utilisant le tableau entier on gagne une quantité supplémentaire égale à Info.

Si f_{ij}= fl_i*fc_j alors on n'a rien gagné par contre plus les f_{ij} sont différents des fl_i*fc_j plus Info est grand. L'étape suivante consiste donc à calculer la contribution $co_{i,j}$ de chaque case du tableau :

$$co_{i,j} = \frac{(f_{i,j} - fl_i\,fc_j)^2}{fl_i\,fc_j},\ i = 1 .. nli,\ j = 1 .. ncol$$

On calcule ensuite le total de l'information apportée par les cases pour exprimer en valeurs relatives le poids de chaque case triée dans l'ordre décroissant.

$$total_info = \sum_{j=1}^{ncol} \left(\sum_{i=1}^{nli} co_{i,j} \right)$$

Le signe de l'écart est calculé par :

$$signe = f_{i,j} - fl_i\,fc_j$$

2. *Exemple numérique.*

Exemple numérique : tableau croisé des codes professionnels des filles et de leur mère 1820-1840

CODE PROFESSIONNEL DES FILLES EN LIGNE CROISE AVEC LE CODE PRO DE LEUR MERE EN COLONNE															
CODE_PRO	Inc	AL	CO	EN	HE	HF	IT	MA	OU	RE	SE	SQ	TE	marges fl(i)	frequences
Inconnu	215	1	15		2	21		2	23	10	18	64	7	378	0.4281
CO	3					1				1				5	0.0057
EN	4			2						1	1			8	0.0091
HE	3				2				2					7	0.0079
HF	38					6	1	1	7	2	13	4		72	0.0815
IT	7								4		2	4		17	0.0193
MA	3								1		2		1	7	0.0079
ME	2		2											4	0.0045
OU	86		3			14			22		10	32		167	0.1891
RE	3								1					4	0.0045

CODE PROFESSIONNEL DES FILLES EN LIGNE CROISE AVEC LE CODE PRO DE LEUR MERE EN COLONNE														
CODE_PRO	Inc	AL	CO	EN	HE	HF	IT	MA	OU	RE	SE	SQ	TE	marges fl(i)	frequences
SE	92		2			8	1	1	9	2	22	37	5	179	0.2027
SQ	20		1						4		2	3	1	31	0.0351
TE	4													4	0.0045
marges fc(j)	480	1	23	2	4	50	2	4	73	16	70	144	14	883	1.0000
fréquences	0.5436	0.0011	0.0260	0.0023	0.0045	0.0566	0.0023	0.0045	0.0827	0.0181	0.0793	0.1631	0.0159	1.0000	

Résultats
Les marges calculées avec le logiciel (vérification avec Excel)

fl := [378, 5, 8, 7, 72, 17, 7, 4, 167, 4, 179, 31, 4]

fc :=[480, 1, 23, 2, 4, 50, 2, 4, 73, 16, 70, 144, 14]

Le tableau des pourcentages de contribution des cases :

	IN	AL	CO	EN	HE	HF	IT	MA	OU	RE	SE	SQ	TE
EIN	.087	.150	.530	-.168	.009	-.002	-.168	.009	-.428	.285	-.939	.018	.033
ECO	.006	-.001	-.026	-.002	-.004	.357	-.002	-.004	-.081	1.794	-.078	-.160	-.016
EEN	-.005	-.002	-.041	42.605	-.007	-.089	-.004	-.007	-.130	.991	.041	-.256	-.025
EHE	-.033	-.002	-.036	-.003	24.013	-.078	-.003	-.006	.686	-.025	-.109	-.224	-.022
EHF	-.007	-.016	-.369	-.032	-.064	.178	.844	.274	.036	.073	1.831	-1.003	-.224
EIT	-.107	-.004	-.087	-.008	-.015	-.189	-.008	-.015	.941	-.061	.062	.107	-.053
EMA	-.033	-.002	-.036	-.003	-.006	-.078	-.003	-.006	.060	-.025	.740	-.224	1.400
EME	-.003	-.001	6.780	-.002	-.004	-.045	-.002	-.004	-.065	-.014	-.062	-.128	-.012
EOU	-.049	-.037	-.082	-.074	-.149	.429	-.074	-.149	.956	-.595	-.156	.164	-.520
ERE	.062	-.001	-.020	-.002	-.004	-.045	-.002	-.004	.266	-.014	-.062	-.128	-.012
ESE	-.057	-.040	-.299	-.080	-.159	-.088	.171	.009	-.447	-.094	.845	.411	.324
ESQ	.116	-.007	.009	-.014	-.028	-.345	-.014	-.028	.158	-.110	-.017	-.164	.103
ETE	.301	-.001	-.020	-.002	-.004	-.045	-.002	-.004	-.065	-.014	-.062	-.128	-.012

Résultat trié du calcul limité à 87 % d’information du tableau, il s’agit d’un cas réel mais assez caricatural :

FILLES ET LEUR MERE

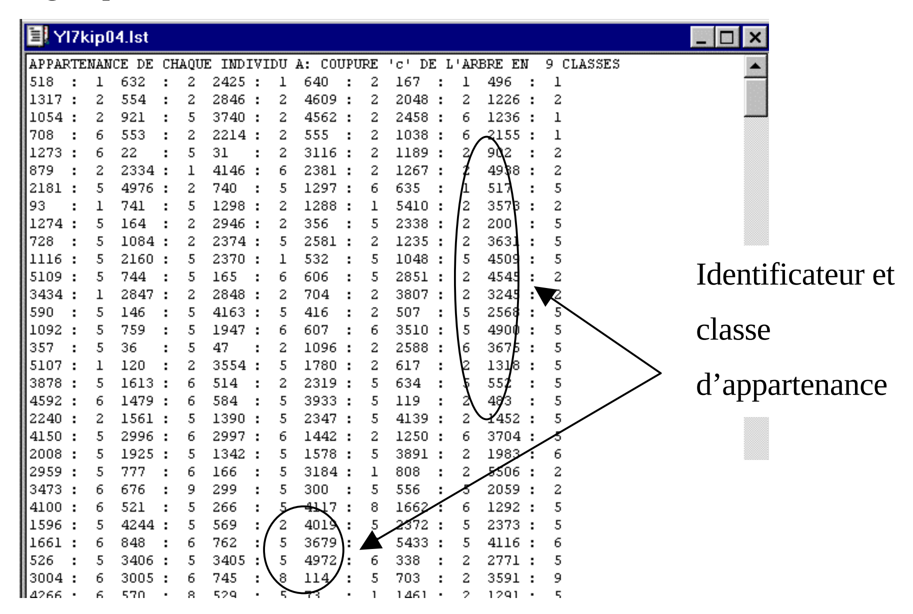
ETAPE	OBJET	Nø	VAR.	Nø	%	SIGNE	% cumul,
1	EEN	3	EN	4	42.61	+	42.61
2	EHE	4	HE	5	24.01	+	66.62
3	EME	8	CO	3	6.78	+	73.40
4	EHF	5	SE	11	1.83	+	75.23
5	ECO	2	RE	10	1.79	+	77.02
6	EMA	7	TE	13	1.40	+	78.42
7	EHF	5	SQ	12	1.00	-	79.43
8	EEN	3	RE	10	0.99	+	80.42
9	EOU	9	OU	9	0.96	+	81.37
10	EIT	6	OU	9	0.94	+	82.31
11	EIN	1	SE	11	0.94	-	83.25
12	ESE	11	SE	11	0.84	+	84.10
13	EHF	5	IT	7	0.84	+	84.94
14	EMA	7	SE	11	0.74	+	85.68
15	EHE	4	OU	9	0.69	+	86.37
16	EOU	9	RE	10	0.59	-	86.96

C. ETAPES DE LA CARTOGRAPHIE

Exemple de la cartographie des effectifs présents aux adresses antérieures, du moment, postérieures pour le groupe IX des 3502 personnes dit des « familles tranquilles ».

1. *Etape 1*

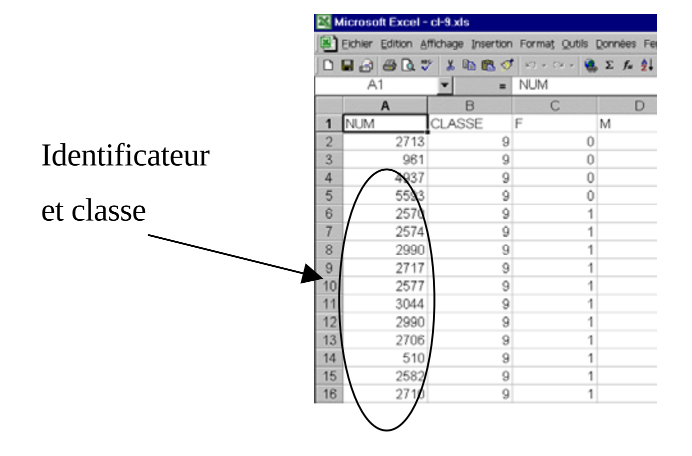
A partir du logiciel Spad d'analyse des données recueil des identificateurs des membres du groupe IX



```
Yl7kip04.lst
APPARTENANCE DE CHAQUE INDIVIDU A: COUPURE 'c' DE L'ARBRE EN  9 CLASSES
518  :  1  632  :  2  2425 :  1  640  :  2  167  :  1  496  :  1
1317 :  2  554  :  2  2846 :  2  4609 :  2  2048 :  2  1226 :  2
1054 :  2  921  :  5  3740 :  2  4562 :  2  2458 :  6  1236 :  1
708  :  6  553  :  2  2214 :  2  555  :  2  1038 :  6  2155 :  1
1273 :  6  22   :  5  31   :  2  3116 :  2  1189 :  2  902  :  2
879  :  2  2334 :  1  4146 :  6  2381 :  2  1267 :  2  4938 :  2
2181 :  5  4976 :  2  740  :  5  1297 :  6  635  :  1  517  :  5
93   :  1  741  :  5  1298 :  2  1288 :  1  5410 :  2  3578 :  2
1274 :  5  164  :  2  2946 :  2  356  :  5  2338 :  2  200  :  5
728  :  5  1084 :  2  2374 :  5  2581 :  2  1235 :  2  3631 :  5
1116 :  5  2160 :  5  2370 :  1  532  :  5  1048 :  5  4509 :  5
5109 :  5  744  :  5  165  :  6  606  :  5  2851 :  2  4545 :  2
3434 :  1  2847 :  2  2848 :  2  704  :  2  3807 :  2  3245 :  2
590  :  5  146  :  5  4163 :  5  416  :  2  507  :  5  2568 :  5
1092 :  5  759  :  5  1947 :  6  607  :  6  3510 :  5  4900 :  5
357  :  5  36   :  5  47   :  2  1096 :  2  2588 :  6  3675 :  5
5107 :  1  120  :  2  3554 :  5  1780 :  2  617  :  2  1318 :  5
3878 :  5  1613 :  6  514  :  2  2319 :  5  634  :  5  552  :  5
4592 :  6  1479 :  6  584  :  5  3933 :  5  119  :  2  483  :  5
2240 :  2  1561 :  5  1390 :  5  2347 :  5  4139 :  2  1452 :  5
4150 :  5  2996 :  6  2997 :  6  1442 :  2  1250 :  6  3704 :  5
2008 :  5  1925 :  5  1342 :  5  1578 :  5  3891 :  2  1983 :  6
2959 :  5  777  :  6  166  :  5  3184 :  1  808  :  2  5506 :  2
3473 :  6  676  :  9  299  :  5  300  :  5  556  :  5  2059 :  2
4100 :  6  521  :  5  266  :  5  4117 :  8  1662 :  6  1292 :  5
1596 :  5  4244 :  5  569  :  2  4019 :  5  2372 :  5  2373 :  5
1661 :  6  848  :  6  762  :  5  3679 :     5433 :  5  4116 :  6
526  :  5  3406 :  5  3405 :  5  4972 :  6  338  :  2  2771 :  5
3004 :  6  3005 :  6  745  :  8  114  :  5  703  :  2  3591 :  9
4266 :  6  570  :  8  529  :  5  73   :  1  1461 :  2  1291 :  5
```


2. *Etape 2*

Nettoyage et mise au format du texte de manière à ce qu'il puisse être relu par Excel en importation.



	A	B	C	D
1	NUM	CLASSE	F	M
2	2713	9	0	
3	961	9	0	
4	4937	9	0	
5	5593	9	0	
6	2570	9	1	
7	2574	9	1	
8	2990	9	1	
9	2717	9	1	
10	2577	9	1	
11	3044	9	1	
12	2990	9	1	
13	2706	9	1	
14	510	9	1	
15	2582	9	1	
16	2710	9	1	

3. *Etape 3*

Intégration de ces données dans Access par import du fichier Excel

Microsoft Access - [Feuil1 Req

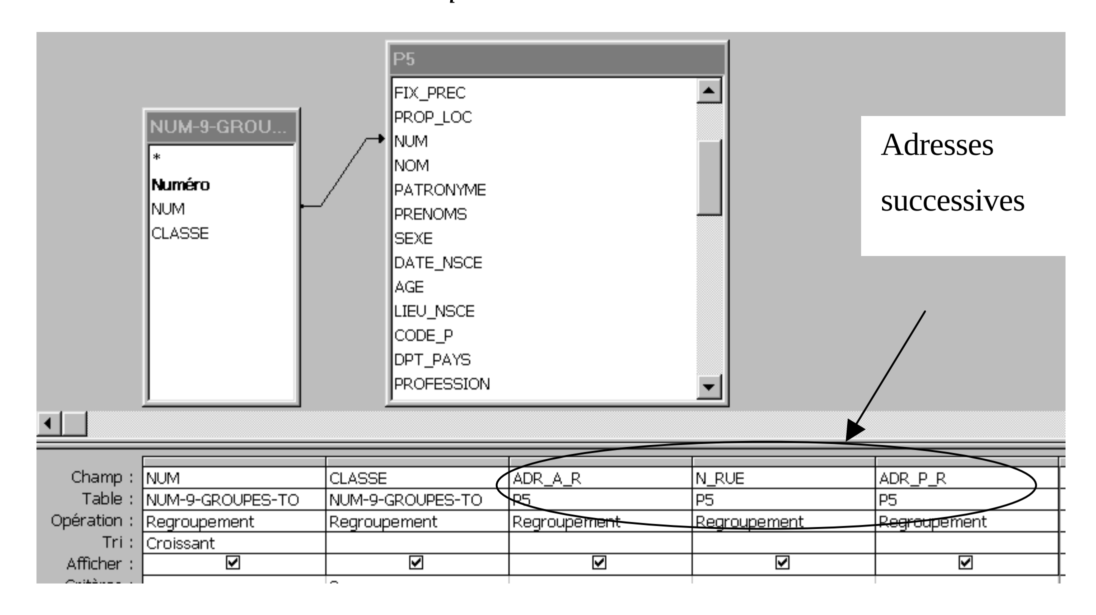
Fichier Edition Affichage Insertio

NUM	CLASSE	F	M	AGE
2713	9	0	0	
961	9	0	0	
4937	9	0	0	
5593	9	0	0	
2570	9	1	0	29
2574	9	1	0	19
2990	9	1	0	49
2717	9	1	0	43
2577	9	1	0	21
3044	9	1	0	46
2990	9	1	0	33
2706	9	1	0	47
510	9	1	0	42
2582	9	1	0	29
2710	9	1	0	38

4. *Etape 4*

Enrichissement par jointure du fichier des identificateurs avec la base des données source P5 de manière à disposer des noms de rue des adresses antérieures, du moment et postérieures dans la même requête. On a maintenant des déclarations.

La requête Access



Le résultat de la requête

	NUM	CLASSE	ADR_A_R	N_RUE	ADR_P_R
	1	9	135	188	188
	1	9	139	188	135
	9	9	81	188	208
	9	9	82	83	81
	9	9	102	188	188
	9	9	188	82	83
	9	9	188	83	101
	9	9	188	102	188
	9	9	188	188	82
	9	9	188	208	103
	9	9	208	103	102
	10	9	81	188	208
	10	9	102	188	188
	10	9	188	82	
	10	9	188	83	101

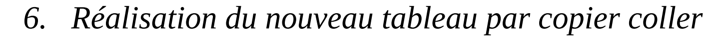
Adresses successives

5. *Exportation et adaptation de ce tableau sous Excel*

C'est nécessaire car Access ne peut pas de manière simple transformer une colonne en ligne. Or pour avoir toutes les adresses il faut qu'à chaque identificateur de personne soit associé ses trois adresses dans la ville en ligne et non pas en colonne.

E36 = 334

	A	B	C	D	E
1	NUM	CLASSE	ADR_A_R	N_RUE	ADR_P_R
2	1	9	135	188	188
3	1	9	139	188	135
4	9	9	81	188	208
5	9	9	82	83	81
6	9	9	102	188	188
7	9	9	188	82	83
8	9	9	188	83	101
9	9	9	188	102	188
10	9	9	188	188	82
11	9	9	188	208	103
12	9	9	208	103	102
13	10	9	81	188	208
14	10	9	102	188	188



6. Réalisation du nouveau tableau par copier coller

	A	B	C
1	NUM	CLASSE	ADR_A_R
2	1	9	135
3	1	9	139
4	9	9	81
8	9	9	188
9	9	9	188
10	9	9	188
11	9	9	188
12	9	9	208
13	NUM	CLASSE	N_RUE
14	1	9	188
15	1	9	188
16	9	9	188
21	9	9	102
22	9	9	188
23	9	9	208
24	9	9	103
25	NUM	CLASSE	ADR_P_R
26	1	9	188
27	1	9	135
28	9	9	208

Fichiers concaténés

Les en-têtes intermédiaires ont été laissés pour la compréhension de la procédure et sont enlevés avant traitement

7. Importation du fichier complet des adresses dans Access pour les tris croisés

Fichier enrichi par jointure avec le champ sexe. Le dernier champ change de nom puisqu'il désigne les trois adresses chronologiquement distinctes

MAISONS-RUES-GROUPE-9 : Table

NUM	CLASSE	SEXE	AA_AM_AP
1	9	F	135
1	9	F	139
1	9	F	188
1	9	F	188
1	9	F	188
1	9	F	135
9	9	F	81
9	9	F	82
9	9	F	102
9	9	F	188
9	9	F	188
9	9	F	188
9	9	F	188
9	9	F	188
9	9	F	208
9	9	F	188
9	9	F	83

8. Résultat du tri croisé Codes des rues X sexe

MAISONS-RUES-GROUPE-9_Analyse croisé

Toutes rues

Effectif

AA_AM_AP	Total	<>	F	M
	1902	12	917	973
1	7		3	4
3	5		5	
5	44		23	21
6	67		46	21
7	9		7	2
10	6		5	1
11	9		6	3
12	12		6	6
13	7		3	4
14	15		10	5
17	9		6	3
20	5		2	3
22	6		2	4
23	34		22	12
24	3		2	1
25	4		3	1
26	55		33	22

9.

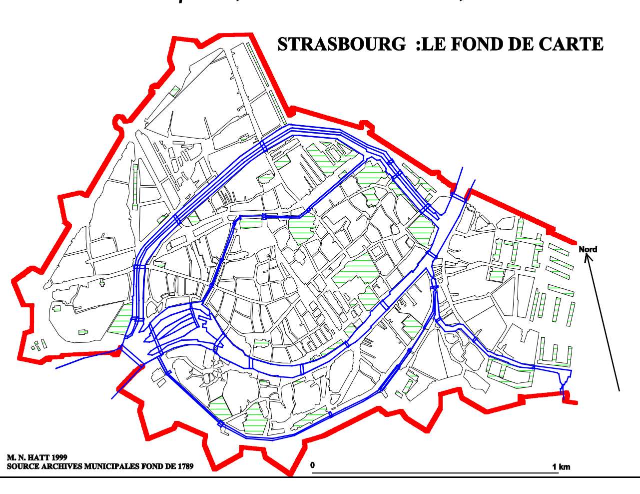
10. Exportation du fichier Access vers un format relu par le logiciel de cartographie, Excel

C’est presque fini mais le logiciel de cartographie ne sait pas relire le format Access donc il faut exporter le tri croisé vers Excel.

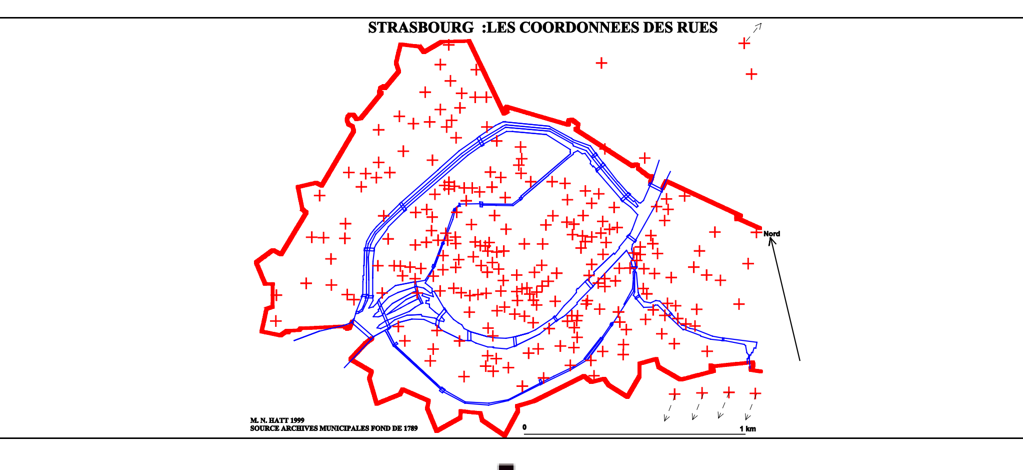
11. La cartographie rassemble trois éléments

Un fond de carte, une localisation en x, y des éléments à cartographier – ici les rues-, associé à un code de repérage, le tableau des données où les identificateurs des rues sont associés aux données à représenter

Le fond, les coordonnées, les données









AA_AM_AP	Situations	<>	F	M
	1902	12	917	973
1	7		3	4
3	5		5	
5	44		23	21
6	67		46	21
7	9		7	2
10	6		5	1
11	9		6	3
12	12		6	6
13	7		3	4
14	15		10	5
17	9		6	3
20	5		2	3
22	6		2	4
23	34		22	12
24	3		2	1
25	4		3	1



carte page suivante

Le résultat : cartographie en cercles proportionnels des effectifs aux adresses antérieures, du moment et postérieures pour le groupe des « familles tranquilles », près de 3000 personnes

ETUDE DE 3502 PERSONNES
ADRESSES DU GROUPE IX 2149 PERSONNES "FAMILLES TRANQUILLES"

LOI RACINE

1 ADR. 575 ADR. 1148 ADR.

Nord

M. N. HATT 1999
SOURCE ARCHIVES MUNICIPALES FOND DE 1789

0 1 km

D. EXEMPLE D'UNE ANALYSE FACTORIELLE SUIVIE D'UNE CLASSIFICATION

Nous avons utilisé Spad version 4 pour réaliser les études statistiques multidimensionnelles du corpus. Pourquoi ce logiciel ? Sa première qualité : ce n'est pas un produit anonyme, il est la création d'Alain Jambu dont les travaux universitaires sont connus depuis longtemps ce qui est un gage de sécurité et de qualité. Il présente d'autre part un certain nombre d'avantages : il est capable d'importer directement des données depuis le tableur en mode ODBC (Open Data Base Connectivity), il permet de visualiser correctement les points doubles sur les graphiques factoriels, il est capable de traiter les données manquantes, il fournit la liste des individus les plus caractéristiques de chaque groupe - les parangons -, enfin, un des points les plus importants, il permet de sauvegarder en fichier la liste des identificateurs des individus appartenant à un groupe de classification ce qui était fondamental pour notre recherche. Statlab, que nos utilisons auparavant (et qui par ailleurs ne semble plus commercialisé), ne le permet pas.

1. Procédure d'analyse du tableau

Les 953 personnes sont déterminées par tri croisé des individus x codes professionnels sous Access.

On procède ensuite à la détermination du nombre de changements professionnels pour chaque individu (entre 1 et 5 changements), ce calcul peut-être mené sous Access ou bien sous Excel.

On élimine ensuite les personnes n'ayant qu'un seul code professionnel. C'est ce tableau final qui est traité (voir en page 4 ci-dessous).

Le tableau des données est traité par analyse[15] en composantes principales normée[16] et classification hiérarchique. Les résultats de l'analyse factorielle

---

[15] Ludovic Lebart, Alain Morineau, Marie Piron, « Statistique exploratoire multidimensionnelle », Dunod, 1995)

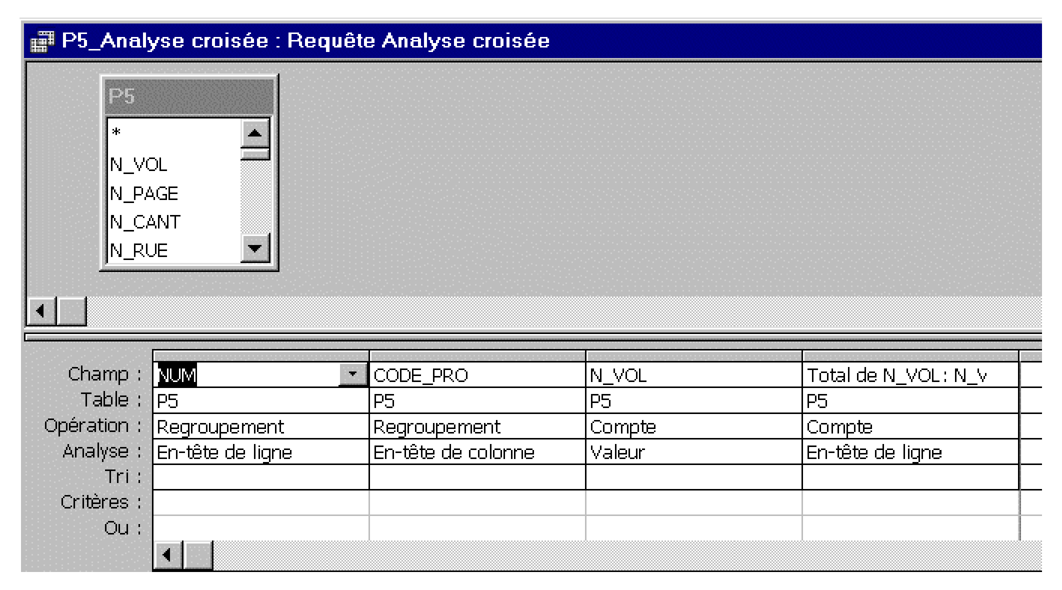
normée sont tout à fait comparables à ceux d'une analyse en correspondances. La classification hiérarchique utilise la caractérisation des d'individus caractérisés par leurs premières coordonnées factorielles.

Etude des groupes complets (moyennes et écart-types pour chaque variable.

Etude et analyse graphique comparative des parangons

---

[16] Dans l'analyse normée chaque colonne du tableau x est traitée de la façon suivante : on commence par calculer la moyenne (xb) et l'écart-type (sigma) de chaque colonne-variable j et pour chaque x_{ij} on calcule ensuite $x_{ij}=x_{ij}$ – xb_j/écart-type$_j$



2. La requête de tri croisé d'Access : personnes x code professionnels

NUM	Total d	<>	AA	AC	AL	AN	AT	BA	CO	CP	DI	EG	EM	EN	ET	HE	HF	IT	MA	ME	MI	OU	PE	PJ	RE	SE	SQ	SV	TB	TE	TM	TR
632	32	3															7		1			5				16						
518	30	1															16									3	10					
167	28	10															4					6				4	4					
2425	28	8															2									18						
554	26	3															18					1				3				1		
640	26	1															7					2				16						
708	23	3															4	10				6										
2048	22	1															6					4				3	8					
3740	21	5																				4				5	7					
921	20			2		1							9	4														4				
2458	18																					4				3	11					
1054	17	17																														
2846	17	5															8					4										
4609	17	4																3				8				2						
1317	16	3																	1							9	3					
1048	16								10		1																	5				

3. Résultat de la requête de tri croisé

M. A. Berlichy, numéro 632 a déclaré 16 fois le code SE, 7 HF, 5 fois OU, 1 fois MA, 3 fois Autre

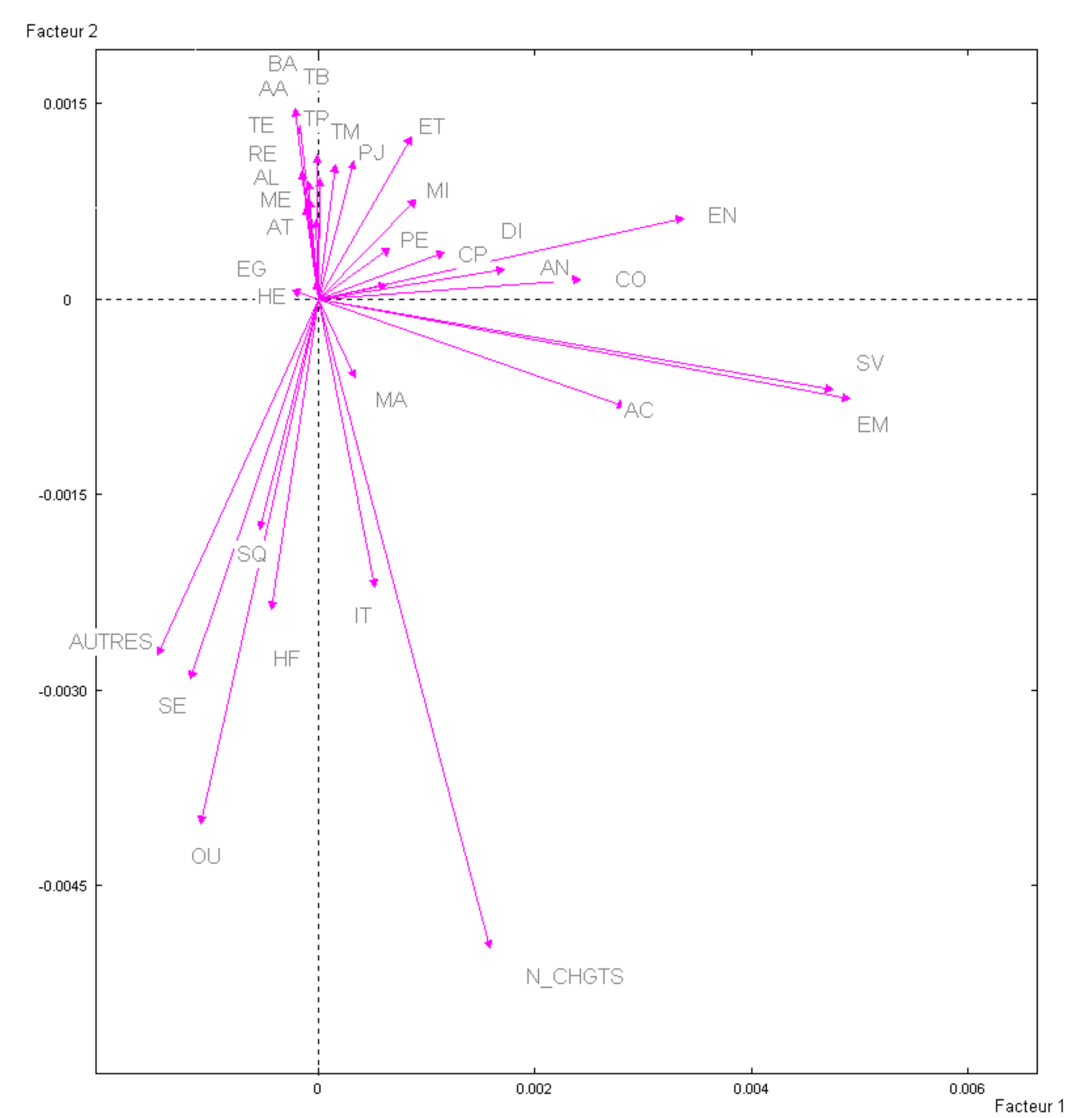
(AU MOINS DEUX CHANGEMENTS) TRI SUR LE NB DE CHGTS DECROISSANT PUIS TOTAL DECL DECROISSANT (2 PAGES EXEMPLE SUR 13)

NNES	TOTAL	Autres	AA	AC	AL	AN	AT	BA	CO	CP	DI	EG	EM	EN	ET	HE	HF	IT	MA	ME	MI	OU	PE	PJ	RE	SE	SQ	SV	TB	TE	TM	TR	CHGTS
Totaux	4767	1123	43	35	114	38	31	42	187	52	26	17	42	45	20	31	419	66	74	17	94	476	51	14	41	779	619	10	107	72	31	51	2177
A0632	32	3															7		1			5				16							5
A0167	28	10															4					6				4	4						5
A0554	26	3															18					1				3				1			5
A2048	22	1															6					4				3	8						5
A0921	20			2		1							9	4														4					5
A0093	13					1			5	5								1									1						5
A2374	11			5					1				3					1									1						5
A2581	10	2															1					2				4	1						5
A0518	30	1															16									3	10						4

NNES	TOTAL	Autres	AA	AC	AL	AN	AT	BA	CO	CP	DI	EG	EM	EN	ET	HE	HF	IT	MA	ME	MI	OU	PE	PJ	RE	SE	SQ	SV	TB	TE	TM	TR	CHGTS
A0640	26							1									7						2				16						4
A0708	23							3									4		10				6										4
A3740	21							5															4				5		7				4
A4609	17							4											3				8				2						4
A1317	16							3													1						9		3				4
A2181	15							6									1										5		3				4

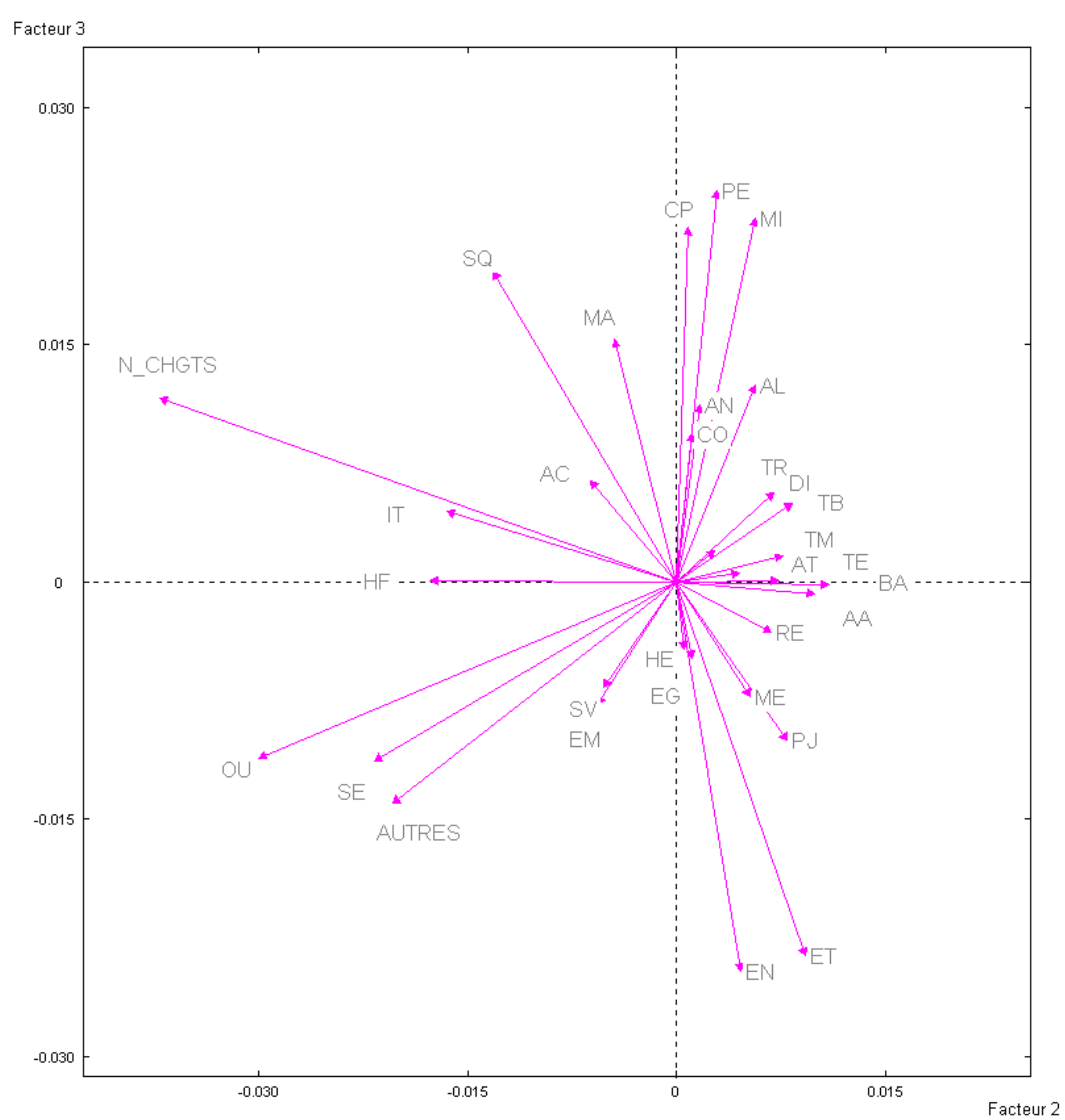
4. Tableau du nombre de changements de codes professionnels 953 personnes x 32 code prof. (avec données manquantes)

Variations professionnelles, plan des facteurs 1 et 2 : 6 % et 5.7 % de la variance



La variance prise en compte par ce premier plan est faible (moins de 12 %). On remarque néanmoins l'opposition forte entre les précaires (SE, SQ, SE, OU, IT) et les autres, AC, SV, EM étant orthogonaux aux premières variables.

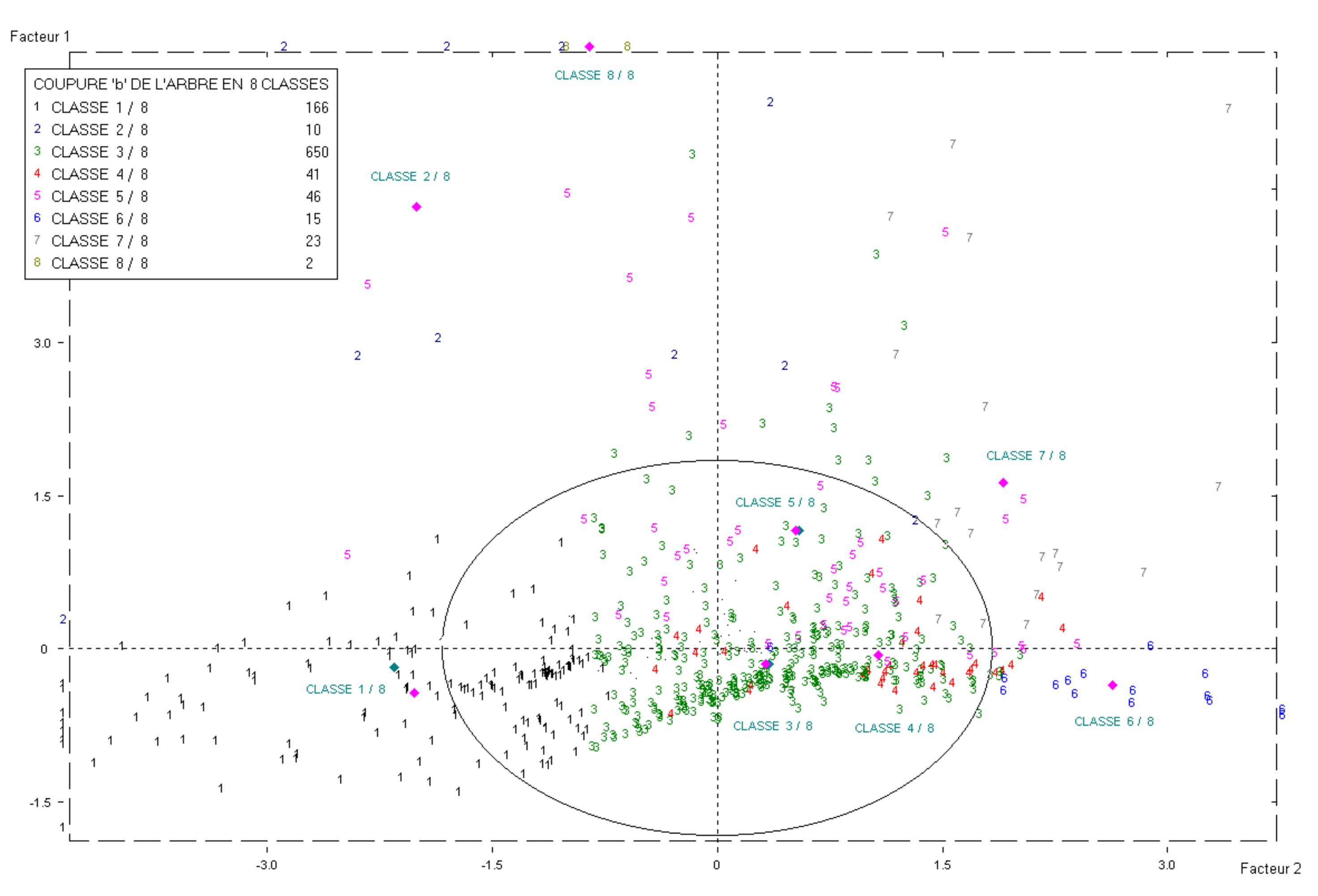
Variations professionnelles, plan des facteurs 2 et 3 : 6 % et 4.4 % de la variance



Dans le deuxième plan factoriel la variable NB_CHGTS tire vers elle toutes les catégories précaires

Variations professionnelles, classification en huit groupes

Zoom sur le groupe central, plan des facteurs 1 et 2 : 6 % et 5.7 % de la variance



Les groupes sont indiqués en couleur par les chiffres du numéro de groupe de 1 à 8, on remarque la très bonne discrimination du groupe I à l'Ouest du graphique, du groupe 6 à l'Est

5. Appartenance de chaque individu aux groupes

Listage de sortie des groupes par Spad. A chaque personne (caractérisée par son numéro) est associé le numéro de groupe. Il suffit de relire ce fichier sous Excel et de le trier pour constituer les groupes. (Une seule page à titre d'exemple)

APPARTENANCE DE CHAQUE INDIVIDU A: COUPURE DE L'ARBRE EN 8 CLASSES

A0632 : 1 A0167 : 1 A0554 : 1 A2048 : 1
A0921 : 8 A0093 : 5 A2374 : 2 A2581 : 1
A0518 : 1 A0640 : 1 A0708 : 2 A3740 : 1
A4609 : 1 A1317 : 1 A2181 : 1 A0047 : 5
A2338 : 1 A0728 : 1 A4146 : 5 A1084 : 1
A2588 : 1 A4116 : 4 A3116 : 1 A1672 : 1
A3679 : 1 A0146 : 1 A1288 : 1 A1452 : 1
A0636 : 1 A0073 : 5 A2129 : 2 A5506 : 1
A3512 : 1 A3391 : 1 A0122 : 1 A3922 : 1
A1933 : 1 A2627 : 1 A2582 : 1 A0210 : 5
A1845 : 1 A2220 : 3 A1215 : 1 A2178 : 1
A3403 : 1 A5419 : 1 A4151 : 1 A4970 : 1
A5158 : 1 A5823 : 1 A4928 : 5 A2425 : 1
A2458 : 1 A2846 : 1 A1947 : 1 A1048 : 8
A1226 : 1 A1038 : 1 A2946 : 1 A3434 : 1
A3807 : 1 A1116 : 1 A2155 : 1 A1297 : 1
A0521 : 5 A4976 : 1 A1189 : 1 A2221 : 1
A0507 : 1 A1267 : 3 A4877 : 2 A2008 : 1
A3753 : 1 A3327 : 5 A3759 : 4 A2847 : 4
A0606 : 5 A0675 : 3 A0634 : 2 A2525 : 1
A2570 : 1 A3118 : 4 A4185 : 1 A3574 : 3
A2809 : 1 A1976 : 1 A0558 : 1 A4799 : 1
A4626 : 1 A4139 : 1 A2045 : 1 A2041 : 1

A1649 : 1 A1859 : 5 A2979 : 1 A2140 : 1

A0570 : 1 A1941 : 1 A2471 : 5 A0556 : 1

A0808 : 1 A2176 : 1 A1590 : 3 A2277 : 1

A2398 : 3 A5109 : 1 A4221 : 1 A3514 : 1

A4508 : 1 A5230 : 1 A4614 : 1 A5029 : 1

A4233 : 1 A4403 : 1 A4643 : 1 A4019 : 1

A1416 : 2 A2728 : 1 A0589 : 1 A2131 : 4

A2852 : 1 A0188 : 1 A0357 : 1 A1370 : 5

A1550 : 1 A0847 : 3 A2065 : 3 A2348 : 1

A1192 : 3 A0753 : 1 A2215 : 1 A2379 : 6

A3639 : 1 A5509 : 4 A4699 : 1 A4020 : 4

A4314 : 1 A2298 : 1 A0901 : 3 A0845 : 1

A3081 : 1 A0663 : 1 A1691 : 5 A1158 : 1

6. Tableau général des moyennes des groupes

Les zéros ont été enlevés pour faciliter la lecture

Code	MOY_Générale	MOY_1	MOY_2	MOY_3	MOY_4	MOY_5	MOY_6	MOY_7	MOY_8
AUTRES	1.18	2.6	0.5	1	0.24	0.3	0.2	0.39	
SE	0.82	2.19		0.63	0.02	0.04			
OU	0.5	1.77	0.6	0.27	0.02	0.04			
SQ	0.65	1.19	0.2	0.49	1.61	0.63	0.13	0.04	
HF	0.44	1.28	0.4	0.31	0.02				
HE	0.03	0.02		0.04					
AA	0.05			0.03		0.02	1.67		
BA	0.04			0.01			2.27		
ME	0.02			0.02			0.07	0.04	
TE	0.08	0.01		0.07			1.53		

Code	MOY_Générale	MOY_1	MOY_2	MOY_3	MOY_4	MOY_5	MOY_6	MOY_7	MOY_8
AL	0.12	0.02		0.04	2.07	0.07			
RE	0.04			0.06					
AT	0.03			0.04		0.07			
EG	0.02	0.02		0.01				0.43	
TB	0.11			0.04	1.88	0.02			
TR	0.05	0.01		0.04	0.54		0.27		
TM	0.03					0.59	0.13		
MA	0.08	0.07	0.3	0.06		0.46			
PJ	0.01							0.52	
IT	0.07	0.16	1.5	0.03		0.04	0.07		
CP	0.05			0.01		0.93			
PE	0.05		0.3	0.02		0.83			
ET	0.02							0.87	
MI	0.1		1	0.12	0.02	0.17			
DI	0.03			0.03	0.02	0.02		0.13	0.5
N_CHGTS	3.01	3	3	2.09	2.2	2.5	2.07	2	4
AN	0.04					0.72	0.07	0.04	0.5
CO	0.2	0.09	0.4	0.14	0.41	0.98		0.22	5
AC	0.04		2.8	0.01		0.02			1
EN	0.05			0.02				1.09	2
SV	0.01					0.02			4.5
EM	0.04		1.2	0.03		0.07		0.04	4.5

7. Variations professionnelles : les parangons

Ce fichier est typiquement difficile à exporter vers le grapheur ou Excel sans un gros travail de nettoyage. Le plus simple est encore d'écrire un filtre de transformation programmé.

```
PARANGONS
CLASSE 1/ 8
EFFECTIF: 166
-----------------------------------------------------------------------------
|RG | DISTANCE | IDENT. ||RG | DISTANCE | IDENT. ||RG | DISTANCE | IDENT. |
+---+----------+--------++---+----------+--------++---+----------+--------+
| 1|0.62348 |A4970 || 2|0.63032 |A1941 || 3|0.65816 |A2852 |
| 4|0.70187 |A5506 || 5|0.71008 |A2277 || 6|0.76242 |A2582 |
| 7|0.76242 |A1845 || 8|0.76806 |A0308 || 9|0.79365 |A2728 |
| 10|0.80041 |A4614 || |  | || |  | |
+---+----------+--------++---+----------+--------++---+----------+--------+
CLASSE 2/ 8
EFFECTIF: 10
-----------------------------------------------------------------------------
|RG | DISTANCE | IDENT. ||RG | DISTANCE | IDENT. ||RG | DISTANCE | IDENT. |
+---+----------+--------++---+----------+--------++---+----------+--------+
| 1| 6.2014 |A1416 || 2| 9.2074 |A0054 || 3| 15.315 |A2129 |
| 4| 31.297 |A2217 || 5| 33.479 |A2329 || 6| 36.901 |A2374 |
| 7| 43.028 |A0634 || 8| 56.079 |A0129 || 9| 64.326 |A4877 |
| 10| 220.22 |A0708 || |  | || |  | |
+---+----------+--------++---+----------+--------++---+----------+--------+
CLASSE 3/ 8
EFFECTIF: 650
-----------------------------------------------------------------------------
|RG | DISTANCE | IDENT. ||RG | DISTANCE | IDENT. ||RG | DISTANCE | IDENT. |
+---+----------+--------++---+----------+--------++---+----------+--------+
| 1|0.88008E-01|A2730 || 2|0.88008E-01|A1289 || 3|0.88008E-01|A0929 |
| 4|0.88008E-01|A5585 || 5|0.88008E-01|A1355 || 6|0.12138 |A1617 |
| 7|0.12138 |A4133 || 8|0.12138 |A4127 || 9|0.12138 |A3227 |
| 10|0.12138 |A4175 || |  | || |  | |
+---+----------+--------++---+----------+--------++---+----------+--------+
CLASSE 4/ 8
EFFECTIF: 41
-----------------------------------------------------------------------------
|RG | DISTANCE | IDENT. ||RG | DISTANCE | IDENT. ||RG | DISTANCE | IDENT. |
+---+----------+--------++---+----------+--------++---+----------+--------+
| 1|0.64074 |A2781 || 2| 1.2471 |A0293 || 3| 1.2471 |A2011 |
| 4| 1.4434 |A3760 || 5| 1.9289 |A1884 || 6| 1.9289 |A2009 |
| 7| 2.1083 |A5509 || 8| 2.2739 |A4020 || 9| 2.2955 |A2131 |
| 10| 2.6075 |A5525 || |  | || |  | |
+---+----------+--------++---+----------+--------++---+----------+--------+
CLASSE 5/ 8
EFFECTIF: 46
-----------------------------------------------------------------------------
|RG | DISTANCE | IDENT. ||RG | DISTANCE | IDENT. ||RG | DISTANCE | IDENT. |
+---+----------+--------++---+----------+--------++---+----------+--------+
| 1| 5.2738 |A2889 || 2| 5.3869 |A4068 || 3| 5.9417 |A4902 |
| 4| 5.9417 |A4445 || 5| 5.9518 |A4247 || 6| 6.3887 |A2056 |
| 7| 6.6113 |A4918 || 8| 7.2601 |A0795 || 9| 8.1135 |A4928 |
| 10| 8.5917 |A0423 || |  | || |  | |
+---+----------+--------++---+----------+--------++---+----------+--------+
```

CLASSE 6/ 8

EFFECTIF: 15

RG	DISTANCE	IDENT.	RG	DISTANCE	IDENT.	RG	DISTANCE	IDENT.
1	0.23487	A4280	2	2.8125	A4638	3	5.2321	A3353
4	5.8395	A4190	5	5.9400	A4688	6	7.6560	A4294
7	8.3454	A1715	8	10.687	A2746	9	11.852	A2379
10	12.053	A3693						

CLASSE 7/ 8

EFFECTIF: 23

RG	DISTANCE	IDENT.	RG	DISTANCE	IDENT.	RG	DISTANCE	IDENT.
1	3.3118	A0831	2	3.3118	A5391	3	3.8540	A2292
4	3.8540	A1492	5	4.2561	A0995	6	5.1790	A3152
7	6.5479	A2199	8	9.8780	A3824	9	12.222	A5456
10	14.420	A1202						

CLASSE 8/ 8

EFFECTIF: 2

RG	DISTANCE	IDENT.	RG	DISTANCE	IDENT.	RG	DISTANCE	IDENT.
1	126.86	A0921	2	126.86	A1048			